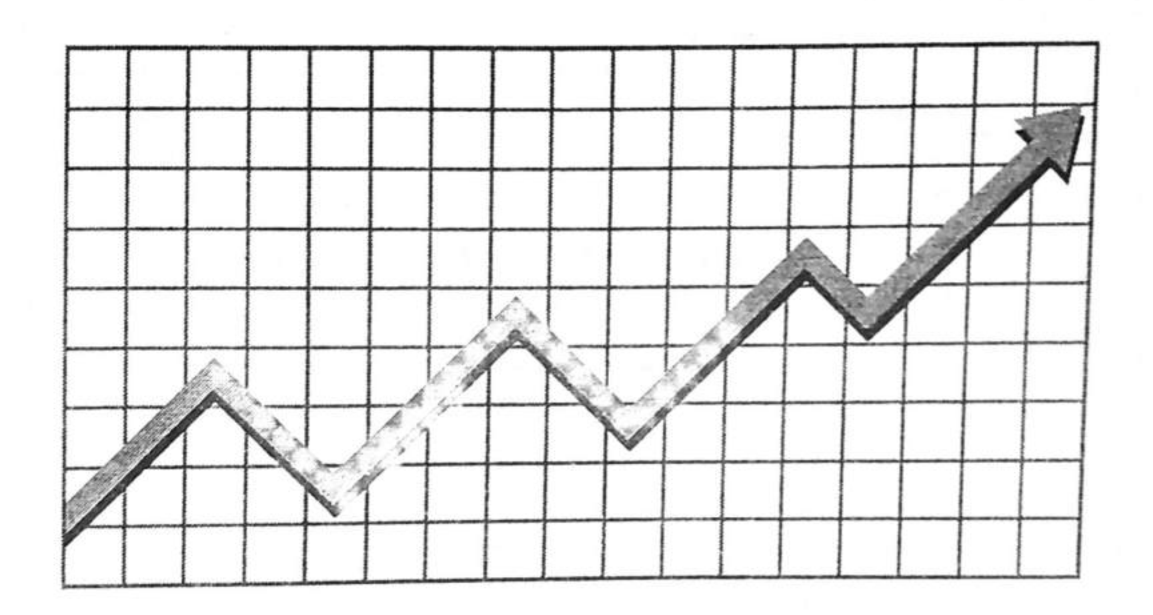
ESTADISTICA Inferencial



Aplicaciones

Segunda edición

Mg. Manuel Córdova Zamora
e-mail: mcordov@pucp.edu.pe

Profesor Principal T.C.
del Departamento de Ciencias de la
Pontificia Universidad Católica del Perú

ESTADÍSTICA: Inferencial
Aplicaciones

Autor: Manuel Córdova Zamora

Nº de Depósito Legal : 1501352002-3878
I.S.B.N. : 9972-813-15-0

Composición y diagramación: Manuel Córdova Zamora

Primera edición: Agosto de 1999
Segunda edición: Febrero del 2006

Obra impresa en los talleres gráficos de:
Distribuidora, Imprenta, Editorial, Librería
MOSHERA S.R.L.
Jr. Tacna 2975 - San Martín de Porres - Lima - Perú
Telefax: 567-9299

Pedidos: Telf. : 534-0638

Impreso en el Perú - Printed in Perú

Presentación a la segunda edición

Esta publicación es la segunda parte revisada y ampliada del libro Estadística descriptiva e Inferencial en su quinta edición. Su contenido cubre los métodos básicos de la estadística inferencial, que regularmente son incluidos en un curso de estadística aplicada.

El objetivo de esta obra es explicar en forma sencilla los métodos de la estadística inferencial, con ejemplos y ejercicios algunos de ellos casos del mundo real, desarrollados con paquetes de computo estadístico o con calculadoras con estadística avanzada, que estimulan la curiosidad científica del lector, conduciéndoles a la formulación de modelos de probabilidad, ANOVA y regresión multivariada, de manera que sirva a su formación básica cualquiera sea su especialidad.

Los ejercicios de este texto han sido resueltos utilizando el paquete de cómputo: Métodos estadísticos básicos, (MCEST), creado por el autor, cuyos resultados son compatibles con los de los paquetes SPSS, ESTADÍSTICA, EXCEL, MINITAB y otros.

Este segundo volumen contiene 9 capítulos. El primero nominado capitulo 8, contiene las distribuciones muestrales. El 9, estimación de parámetos. El 10, prueba de hipótesis. El 11, pruebas con chi-cuadrado. El 12, análisis de varianza. El 13, regresión simple y multivariada. El 14 introducción a las series de tiempo. El 15, una introducción a las pruebas no paramétricas. Y el 16, una introducción a la teoría de la decisión estadística.

Estoy muy agradecido por la acogida que recibe esta publicación, y me motiva entonces, a realizar una revisión permanente del texto ampliando sus aplicaciones. Quiero expresar también mi agradecimiento a la Pontificia Universidad Católica del Perú por permitirme realizar este trabajo fruto de mi experiencia en sus aulas. A mis alumnos de la facultad Ciencias Administrativas y Contabilidad. A mis alumnos de estudios generales ciencias que aplican conmigo los tres primeros capítulos de este texto. Así mismo, expreso mi agradecimiento a los colegas de Lima y provincias por la aplicación del texto a cursos de estadística inferencial o estadística aplicada a diversas ramas del saber.

Lima, setiembre del 2002
Manuel Córdova Zamora

CONTENIDO

Capítulo 8

DISTRIBUCIONES MUESTRALES

8.1 Muestreo aleatorio

8.1.1 Población y parámetros

Definición. Se denomina **población** o universo a la totalidad de personas u objetos que tienen una o más características medibles o contables de naturaleza cualitativa o cuantitativa.

La característica medible o contable es una *variable estadística* cuyo valor, numérico o no numérico, es una *observación*.

Si la variable estadística a estudiar es una sola, cada elemento de la población puede asociarse con una observación. En este sentido, se denomina población al conjunto de valores posibles de la variable.

Si los elementos de la población se definen en forma aleatoria, entonces la variable estadística cuantitativa es una variable aleatoria cuyos valores constituyen la población. En este caso, la distribución de la población es la distribución de la variable aleatoria, por lo tanto, la media y la varianza de la variable aleatoria, vienen a ser la media y la varianza de la población.

Si la variable aleatoria X tiene distribución $f(x)$, se puede referir a la población $f(x)$. Por ejemplo, si X está normalmente distribuida se dice que la población está normalmente distribuida o que se tiene una población normal.

Por el número de observaciones la población puede ser *finita* de tamaño N, o *infinita*. Algunas poblaciones finitas son tan grandes que en teoría son asumidas como poblaciones infinitas.

Definición. Se denominan **parámetros** a las medidas descriptivas que caracterizan a la distribución de la población. Entre otros, los parámetros poblacionales son:

Media : μ
Proporción : π o p
Varianza : σ^2
Desviación estándar: σ

En diversas aplicaciones estadísticas al estudiar una población, la variable aleatoria que la define puede tener distribución conocida o no. La distribución de la población es conocida, si se conocen sus parámetros y su forma, es decir si se conoce su distribución de probabilidad.

Si la distribución de la población es desconocida, podemos estar interesados en:

* **Estimar sus parámetros**, si se conoce su distribución, y

* **Probar determinada suposición** acerca de un valor determinado del parámetro, o probar la suposición acerca del tipo de distribución de probabilidades de la población.

8.1.2 Muestra aleatoria

En vez de examinar la población entera, lo cual puede resultar físicamente imposible o no práctica, puede examinarse una **muestra** de la población con el propósito de **inferir** los resultados encontrados.

Una muestra es un subconjunto de la población.

El proceso de selección de una muestra de n elementos de la población se llama **muestreo**. Las ventajas y las razones para el muestreo son diversas, las mismas que no explicaremos en este texto.

El proceso que consiste en inferir resultados a la población a partir de la muestra se denomina **inferencia estadística.** La confiabilidad de las conclusiones extraídas concernientes a una población depende de sí la muestra se ha escogido apropiadamente de manera que represente bien a la población.

En general existen dos tipos de muestras: Las **no probabilísticas** (basadas en el criterio de expertos) y las **probabilísticas**

A las muestras probabilísticas se les llama también **muestras aleatorias**..

Se llama muestreo aleatorio a todo proceso que asegure en cualquier momento del mismo igual probabilidad (distinta de cero) de ser incluidos en la muestra a todos los elementos que pertenezcan a la población en dicho momento.

Las muestras aleatorias son de 4 tipos: Al azar simple, al azar sistemático, estratificado y por grupos (o conglomerados).

Muestra al azar simple

Es aquella en la que los elementos de la muestra se escogen del total de la población en forma individual con una oportunidad igual e independiente. Por lo general se utiliza una tabla de números aleatorios o un programa de computo generador de números aleatorios para identificar a los elementos numerados de la población que se eligen para la muestra.

Si la población es infinita el muestreo aleatorio ocurre cuando la extracción de los elementos de la muestra se hace con o sin reemplazo. Si la población es finita de tamaño N, el muestreo aleatorio ocurre también si la extracción es con o sin reemplazo.

Con reemplazo, la probabilidad de cada elemento de ser extraído es $1/N$.

Si es sin reemplazo, la probabilidad de cada elemento de ser elegido es $1/N$ en la primera extracción, es de $1/(N-1)$ en la segunda extracción, es $1/(N-2)$ en la tercera extracción, etc.

Seleccionar una muestra al azar simple es similar a la que se realiza en la extracción aleatoria de números en una lotería.

Un modo más conveniente de seleccionar muestras al azar simple es enumerar a todos los elementos de la población y luego usar una **tabla de números aleatorios**. (Ver por ejemplo página 223 y apéndice *E* de la referencia 8)

Por **ejemplo**, si queremos seleccionar una muestra al azar simple de 4 alumnos de la lista de la clase que tiene 50 alumnos, se escriben los números 01 a 50, se colocan en una urna, se chocolatea y luego se escogen 4 de las fichas de la urna. El lector debería hacer esta selección con una tabla de números al azar.

Muestra al azar sistemática

Una muestra aleatoria sistemática es aquella en que sus elementos se eligen de la población a intervalos uniformes a partir de un listado ordenado. El k-ésimo elemento de la muestra es $k=N/n$, donde n es el tamaño de la muestra y N el tamaño de la población.

Por **ejemplo**, al elegir una muestra sistemática de 100 alumnos de los 3000 alumnos que tiene Estudios Generales ciencias de la PUCP, $k = 3000/100 = 30$. El primero se elige en forma aleatoria de los 30 primeros de la lista y los demás sistemáticamente cada 30 alumnos de la lista.

Muestreo aleatorio estratificado

Primero se clasifican a los elementos de la población en subgrupos separados de acuerdo con una o más características importantes (estratos). Después se obtiene por separado una muestra aleatoria simple o sistemática en cada estrato.

El tamaño de cada submuestra debe ser proporcional al tamaño del estrato para asegurar representatividad.

Por **ejemplo**, para obtener una muestra aleatoria de 600 electores de una población de 600,000 electores de los cuales 300,0000 son de clase baja, 200,000 de clase media y 100,000 de clase alta. Se deben elegir al azar 300 de clase baja, 200 de clase media y 100 de clase alta.

Muestreo aleatorio por conglomerados

Denominado también muestreo agrupado se utiliza cuando se trata de obtener una muestra al azar de una población dispersa en una gran área geográfica (ver referencia 8 página 227).

Los elementos de la población se dividen en forma natural en subgrupos. Luego se eligen al azar los subgrupos que forman la muestra.

Por **ejemplo**, al estudiar los pensiones que se pagan en los colegios particulares de Lima, sería difícil obtener una lista de todas las pensiones que forman la población, pero puede obtenerse una lista de los colegios particulares de Lima (grupos). Entonces, con esta lista puede obtener una muestra aleatoria de colegios y así obtener los pensiones que se pagan en estos colegios.

El **muestreo aleatorio simple**, es pues el proceso de selección de una muestra por el cual cada uno de los elementos de la población tienen una oportunidad igual e independiente de ser incluidos en la muestra. En el muestreo aleatorio simple cada variable aleatoria X_i cuyo valor es x_i, tiene la misma distribución de la población de la cual se obtiene.

Por **ejemplo**, supongamos que una población consiste de 8 fichas, dos con el número 2, cuatro con el número 5, y dos con el número 7. Si se extrae una ficha al azar, la ficha puede tomar cualquiera de los tres valores: 2 con probabilidad 0.25, 5 con probabilidad 0.50, y 7 con probabilidad 0.25, que viene a ser la misma distribución de la población.

Luego, diremos que los valores $x_1, x_2, \ldots x_n$ tomados respectivamente por las variables aleatorias $X_1, X_2, \ldots, X_n$, constituyen una muestra aleatoria simple de tamaño n de una población $f(x)$ de la variable aleatoria X, si estas variables aleatorias están distribuidas en forma idéntica a la distribución de la población y son independientes.

Llamaremos también muestra aleatoria simple a este conjunto de variables aleatorias. Formalmente definimos una muestra aleatoria simple o brevemente muestra aleatoria de la forma siguiente:

Definición. (Muestra aleatoria simple). Dada una población $f(x)$ con media μ y varianza σ^2, se denomina muestra aleatoria de tamaño n de esa población, a un conjunto de n variables aleatorias $X_1, X_2, ..., X_n$ tales que:

1) Son independientes.

2) Cada una de ellas está distribuida en forma idéntica a $f(x)$.

La condición 1) implica que la **distribución de probabilidad conjunta** de la muestra aleatoria $X_1, X_2, ..., X_n$ es la expresión:

$$f(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n)$$

La condición 2) significa que:

a) Cada variable aleatoria X_i tiene la misma media y varianza de la distribución de X, es decir: $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$.

b) La distribución de probabilidad de cada variable aleatoria X_i es la misma distribución de probabilidades de X, esto es, $f(x_i) = f(x)$.

NOTA. El proceso de obtener este tipo de muestra requiere población infinita o bien población finita pero con reposición de elementos.

EJEMPLO 8.1.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de una población normal $N(\mu, \sigma^2)$.

a) Escriba la función de densidad de probabilidad conjunta de la muestra.

b) Si $n = 6$, $\mu = 20$, y $\sigma^2 = 25$, calcule la probabilidad de que:

b1) $X_1 + X_3 + X_4 - X_6$ sea mayor que 52.

b2) al menos una de las X_i sea menor que 29.8.

SOLUCION.

a) La función de densidad conjunta de la muestra aleatoria es

$$f(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n) = [f(x_i)]^n$$

$$f(x_1, x_2, \ldots x_n) = \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \right]^n = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

b1) La media y la varianza de la variable aleatoria $Y = X_1 + X_3 + X_4 - X_6$ están dadas respectivamente por:

$$E(Y) = E(X_1) + E(X_3) + E(X_4) - E(X_6) = 20 + 20 + 20 - 20 = 40.$$

$$V(Y) = V(X_1) + V(X_3) + V(X_4) + V(X_6) = 25 + 25 + 25 + 25 = 100.$$

Por la propiedad reproductiva de la normal la variable aleatoria Y tiene distribución normal $N(40,100)$, luego, la variable aleatoria estándar:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} = \frac{Y - 40}{10}, \quad \text{tiene distribución} \quad N(0,1), \quad \text{y}$$

$$P[Y > 52] = P\left[\frac{Y-40}{10} > \frac{52-40}{10}\right] = P[Z > 1.2] = 0.1151.$$

b2) Sea ahora la variable aleatoria

$$Y_i = \begin{cases} 1, & \text{si } X_i < 29.8 \\ 0, & \text{si } X_i \geq 29.8 \end{cases},$$

Entonces, Y_i es Bernoullí $B(1, p)$, donde $p = P[Y_i = 1]$, la probabilidad del éxito es igual a:

$$p = P[X_i < 29.8] = P\left[\frac{X_i - 20}{5} < \frac{29.8 - 20}{5}\right] = P[Z < 1.96] = 0.975.$$

En consecuencia, la variable aleatoria: $Y = \sum_{i=1}^{6} Y_i$ es binomial $B(6, p)$, esto es,

$$P[Y = y] = C_y^6 p^y (1-p)^{6-y}, \quad y = 0, 1, 2, 3, 4, 5, 6,$$

Por tanto, la probabilidad de que al menos un X_i, sea menor que 29.8 es:

$$P[Y \geq 1] = 1 - P[Y = 0] = 1 - (0.025)^6 = 1 - 0.000 = 1.000.$$

8.1.3 Estadísticas

Definición. Se denomina **estadística** a cualquier función de las variables aleatorias que constituyen la muestra.

Una estadística es pues una variable aleatoria $Y = H(X_1, X_2, \ldots, X_n)$, cuyo valor es el número real $y = H(x_1, x_2, \ldots x_n)$.

El término estadística se usa para referirse tanto a la función de la muestra o **variable aleatoria,** como al **valor** de esta variable.

En general para cada parámetro poblacional hay una estadística correspondiente a calcularse a partir de la muestra.

Algunas estadísticas importantes y sus valores calculados a partir de una muestra aleatoria de tamaño *n* son:

a) La media muestral: $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (variable aleatoria), $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$. (valor)

b) La varianza muestral: $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, con valor: $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

c) La desviación estándar muestral: $S = \sqrt{S^2}$, con valor: $s = \sqrt{s^2}$

d) La proporción muestral: $\hat{P}$ o $\overline{P} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (Porcentaje de éxitos en la muestra)

donde $X_i \sim B(1, p)$ (el parámetro *p* es el porcentaje de éxitos en la población).

También,

$$\overline{P} = \frac{X}{n}, \quad \text{donde } X \sim B(n, p)$$

El valor de $\overline{P}$ (o $\hat{P}$), calculada a partir de una muestra es $\bar{p}$ (o $\hat{p}$) $= x/n$

NOTA. Error de muestreo es la diferencia entre una estadística de la muestra y el parámetro correspondiente **de la población.**

NOTA. Una estadística importante también es el total de la muestra $\sum_{i=1}^{n} X_i = n\overline{X}$.

Su parámetro correspondiente es **el total de la población finita.**

8.2 Distribuciones muestrales

Definición. Se denomina **distribución muestral** de una estadística a la distribución de probabilidad. de esa variable aleatoria

Por ejemplo, a la distribución de probabilidad de la estadística media: $\overline{X}$, se le denomina distribución muestral de la media.

Las aplicaciones de las distribuciones muestrales son aplicaciones del **teorema central del límite**

8.2 .1 Distribución muestral de la media $\overline{X}$

TEOREMA 8.1. Sea $X_1, X_2, ..., X_n$, una muestra aleatoria de tamaño n escogida de una población $f(x)$ que tiene media μ y varianza σ^2.
Si $\overline{X}$ es la media muestral , entonces,

a) $E(\overline{X}) = \mu$ b) $Var(\overline{X}) = \dfrac{\sigma^2}{n}$

c) Para n suficientemente grande, la variable aleatoria,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

tiene distribución aproximadamente normal $N(0,1)$.

PRUEBA.

Por la definición de muestra aleatoria, las variables aleatorias $X_1, X_2, ..., X_n$. son independientes e idénticamente distribuidas como $f(x)$ con $E(X_i) = \mu$, y con $Var(X_i) = \sigma^2$. Entonces,

a) $E(\overline{X}) = E\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n} n\mu = \mu.$

b) $Var(\overline{X}) = V\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} V(X_i) = \dfrac{1}{n^2} n\sigma^2 = \dfrac{\sigma^2}{n}.$

c) Se deduce del **teorema del límite central** escribiendo

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}.$$

NOTAS.

1. La aproximación de $\overline{X}$ a la normal $N(\mu, \sigma^2/n)$ es buena si $n \geq 30$, sin importar si la población es discreta o continua.

2. Si la muestra aleatoria es escogida de una población normal $N(\mu, \sigma^2)$, entonces, la distribución de $\overline{X}$ es exactamente normal $N(\mu, \sigma^2/n)$, para cualquier tamaño de muestra, $n \geq 2$.

3. La varianza de la media: $Var(\overline{X}) = \dfrac{\sigma^2}{n}$ es válida, si el ***muestreo es con o sin reemplazo*** en una población infinita, (o es con reemplazo en una población finita de tamaño *N*).
   Si **el muestreo es sin reemplazo en una población finita de tamaño** *N*, entonces, la varianza de la distribución de $\overline{X}$ es:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

   El coeficiente $\dfrac{N-n}{N-1}$ se denomina **factor de corrección para población finita**.
   Observe que cuando $N \to +\infty$ el factor de corrección tiende a uno,

4. La desviación estándar de una estadística es conocida como **error estándar**.

EJEMPLO 8.2.

Una fábrica textil tiene 5 operarios. Los años de servicio en la fábrica de estos operarios son

$$3, \ 4, \ 7, \ 9, \ 12.$$

a) Calcule la media y la varianza de la población de años de servicios.
b) Determine la distribución de la media de las muestras de tamaño dos escogidas de la población (sin reposición).
c) Determine la distribución de la media de las muestras de tamaño dos escogidas con reposición.
d) Si se extraen muestras al azar de tamaño 36 con reposición, ¿cuál es la probabilidad de que la media muestral esté entre los valores 5 y 8?

SOLUCION

a) La distribución de probabilidad de esta población finita de tamaño N=5, es la distribución uniforme siguiente:

x_i	3	4	7	9	12
$f(x_i) = P[X = x_i]$	1/5	1/5	1/5	1/5	1/5

La media y la varianza de la población son respectivamente:

$$\mu = \sum_{i=1}^{N} x_i f(x_i) = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{3+4+7+9+12}{5} = 7.$$

$$\sigma^2 = \sum_{i=1}^{N} x_i^2 f(x_i) - \mu^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2 = \frac{3^2+4^2+7^2+9^2+12^2}{5} - 7^2 = 10.8.$$

b) Se pueden extraer $C_2^5 = 5\times 4 = 20$ muestras de tamaño 2 (sin reposición)

Las muestras y sus medias respectivas son las siguientes:

Muestras				**Medias de las muestras**			
3, 4	3, 7	3, 9	3, 12	3.5	5	6	7.5
4, 3	4, 7	4, 9	4, 12	3.5	5.5	6.5	8
7, 3	7, 4	7, 9	7, 12	5	5.5	8	9.5
9, 3	9, 4	9, 7	9, 12	6	6.5	8	10.5
12, 3	12, 4	12, 7	12, 9	7.5	8	9.5	10.5

La distribución de probabilidades de la media es:

$\overline{x}$	3.5	5	5.5	6	6.5	7.5	8	9.5	10.5
$f(\overline{x})$	2/20	2/20	2/20	2/20	2/20	2/20	4/20	2/20	2/20

Luego, $\mu_{\overline{X}} = E(\overline{X}) = \sum f(\overline{x})\overline{x} = 140/20 = 7.$

$$\sigma_{\overline{X}}^2 = Var(\overline{X}) = \sum f(\overline{x})\overline{x}^2 - \mu^2 = \frac{1061}{20} - 7^2 = 4.05.$$

Observe también que:

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{10.8}{2}\left(\frac{5-2}{5-1}\right) = 4.05\,.$$

C) Se pueden extraer $5\times 5=25$ muestras de tamaño dos con reposición. Las muestras y sus medias son las siguientes.

Muestras					Medias de las muestras				
3 ,3	3, 4	3, 7	3, 9	3, 12	3	3.5	5	6	7.5
4 ,3	4, 4	4, 7	4, 9	4, 12	3.5	4	5.5	6.5	8
7 ,3	7 ,4	7 ,7	7 ,9	7 ,12	5	5.5	7	8	9.5
9 ,3	9, 4	9, 7	9, 9	9, 12	6	6.5	8	9	10.5
12 ,3	12, 4	12, 7	12, 9	12, 12	7.5	8	9.5	10.5	12

La distribución de probabilidades de las medias es:

$\bar{x}$	3	3.5	4	5	5.5	6	6.5	7	7.5	8	9	9.5	10.5	12
$f(\bar{x})$	1/25	2/25	1/25	2/25	2/25	2/25	2/25	1/25	2/25	4/25	1/25	2/25	2/25	1/25

Luego,

$$\mu_{\bar{X}} = E(\bar{X}) = \sum f(\bar{x})\bar{x} = 175/25 = 7.$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \sum f(\bar{x})\bar{x}^2 - \mu^2 = \frac{1360}{25} - 7^2 = 5.4.$$

Observar también que : $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n} = \dfrac{10.8}{2} = 5.4.$

d) Sea $\bar{X}$ la media de las muestras de tamaño $n=36$ con reposición.

La estadística $\bar{X}$ tiene media y varianza respectivas:

$\mu_{\bar{X}}$ o $E(\bar{X}) = \mu_X = 7,$

$\sigma^2_{\bar{X}}$ o $Var(\bar{X}) = \dfrac{\sigma^2}{n} = \dfrac{10.8}{36} = 0.3$

El **error estándar** de $\bar{X}$, es $\sigma_{\bar{X}} = \sqrt{0.3} = 0.55$

Entonces, la variable estándar,

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 7}{0.55}$$

tiene distribución aproximadamente normal $N(0,1)$. Por tanto,

$$P[5 \le \bar{X} \le 8] = P\left[\frac{5-7}{0.55} \le \frac{\bar{X}-7}{0.55} \le \frac{8-7}{0.55}\right] = P[-3.64 \le Z \le 1.82] = 0.9655.$$

EJEMPLO 8.3.

El número de automóviles por familia es una variable aleatoria X cuya distribución de probabilidad es como sigue:

x	0	1	2	3	4
$f(x)$	4/12	4/12	2/12	1/12	1/12

a) Halle la media y la varianza de la población del número de automóviles por familia

b) Si se escoge al azar una muestra de 49 familias, ¿cuál es la probabilidad de que la media muestral de autos por familia esté entre 1 y 2?.

SOLUCION.

a) La media y la varianza de X son respectivamente:

$$\mu_X = E(X) = \sum x_i f(x_i) = 0(\frac{4}{12}) + 1(\frac{4}{12}) + 2(\frac{2}{12}) + 3(\frac{1}{12}) + 4(\frac{1}{12}) = 1.25.$$

$$\sigma_X^2 = E(X^2) - (\mu_X)^2 = \sum x_i^2 f(x_i) - (\mu_X)^2 = 3.08 - (1.25)^2 = 1.52.$$

b) Sea $\overline{X}$ la media del número de automóviles en muestras de 49 familias. La estadística $\overline{X}$ tiene distribución aproximadamente normal con media y varianza respectivas:

$$\mu_{\overline{X}} \text{ o } E(\overline{X}) = \mu_X = 1.25$$

$$\sigma_{\overline{X}}^2 \text{ o } Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{1.52}{49} = 0.031.$$

El error estándar de la media $\overline{X}$ es:

$$\sigma_{\overline{X}} = \sqrt{0.031} = 0.176$$

Entonces, la variable estándar,

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 1.25}{0.176}$$

tiene distribución aproximadamente normal $N(0,1)$. Por tanto,

$$P[1 \le \overline{X} \le 2] \cong P\left[\frac{1-1.25}{0.176} \le Z \le \frac{2-1.25}{0.176}\right] \cong P[-1.42 \le Z \le 4.26] = 0.9222.$$

EJEMPLO 8.4

Un auditor toma una muestra aleatoria de 100 cuentas por cobrar de una población de 500 cuentas por cobrar. El auditor sabe que las 500 cuentas por cobrar constituyen una población finita cuya desviación estándar es $\sigma = \$145$. ¿Cuál es la probabilidad de que la media muestral difiera de la media poblacional en mas de \$26?

SOLUCION.

Sea $\overline{X}$ la media de la muestra de tamaño $n = 100$ escogida de la población finita de $N = 500$ casos. Entonces, la variable aleatoria $\overline{X}$ tiene distribución aproximadamente normal con media $\mu_{\overline{X}} = \mu$ y error estándar:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{145}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} = 12.982$$

En consecuencia, la variable aleatoria estándar:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{12.982}$$

tiene distribución normal $N(0,1)$. La probabilidad de que la media muestral difiera de la media poblacional en más de \$26 es:

$$P\left[\left|\overline{X} - \mu\right| > 26\right] = P\left[|Z| > \frac{26}{12.982}\right] = P\left[|Z| > 2.00\right] = 0.0456$$

EJEMPLO 8.5.

Se puede ajustar una máquina de refrescos de tal manera que llene los vasos con un promedio de μ onzas por vaso, donde la cantidad de onzas por vaso tiene una distribución normal con una desviación estándar de 0.5 onzas

a) Encuentre el valor de μ de tal manera que al llenar vasos de 10 onzas solamente se derramen el 3% de los vasos

b) Con el valor de μ hallado en a) encuentre la probabilidad de que al llenar 100 vasos de 10 onzas el promedio del líquido derramado sea mayor de 0.06 onzas.

c) ¿Con cuántos vasos de 10 onzas se consigue que el contenido promedio del liquido sea menor al promedio de la población en 0.1225 onzas con probabilidad igual a 0.025?.

SOLUCION.

Sea X el contenido en vasos de 10 onzas que llena la máquina de refrescos. La distribución de X es normal, esto es, $X \sim N(\mu, 0.5^2)$.

a) Se debe calcular μ tal que $P[X > 10] = 0.03$

$$0.03 = P[X > 10] = P\left[Z > \frac{10-\mu}{0.5}\right]$$

de donde resulta:

$$\frac{10-\mu}{0.5} = 1.88, \quad \text{y} \quad \mu = 9.06$$

b) Sea $\overline{X}$ la media de los contenidos de 100 vasos. Por el teorema del límite central, la variable aleatoria $\overline{X}$ tiene distribución aproximadamente normal con media $\mu_{\overline{X}} = 9.06$ onzas y error estándar:

$$\sigma_{\overline{X}} = \sigma / \sqrt{n} = 0.5 / \sqrt{100} = 0.05.$$

Entonces, la variable aleatoria estándar:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 9.06}{0.05}$$

tiene distribución normal $N(0,1)$.

La probabilidad de que al llenar 100 vasos de 10 onzas el promedio del líquido derramado sea mayor de 0.06 onzas es:

$$P[\overline{X} > 10.06] = P\left[Z > \frac{10.06 - 9.06}{0.05}\right] = P[Z > 20] = 0.000$$

c) Sea $\overline{X}_n$ la media de la muestra de tamaño n. Entonces, la distribución de $\overline{X}_n$ es normal con media igual a 9.06, y error estándar igual a $0.5/\sqrt{n}$. Se debe calcular n tal que $P[\overline{X}_n < 8.9375] = 0.0250$. Entonces,

$$0.0250 = P[\overline{X}_n < 8.9375] = P\left[Z < \frac{8.9375 - 9.06}{0.5/\sqrt{n}}\right].$$

De donde resulta,

$$\frac{-0.1225}{0.5}\sqrt{n} = -1.96 \qquad \sqrt{n} = 8$$

$$n = 64$$

8.2.2 Distribución muestral de la proporción

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n escogida de la población de Bernoullí $B(1, p)$, donde p es el porcentaje de éxitos en la población y sea

$$\overline{P} = \frac{X_1 + X_2 + ...X_n}{n} = \frac{X}{n}$$

la **proporción de éxitos en la muestra,** siendo, $X = X_1 + X_2 + \cdots + X_n$ una variable binomial $B(n, p)$, entonces,

a) $\mu_{\overline{P}} = E(\overline{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p.$

b) $\sigma^2_{\overline{P}} = V(\overline{P}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}[np(1-p)] = \frac{p(1-p)}{n}.$

c) Si n es suficientemente grande, entonces la variable aleatoria:

$$Z = \frac{\overline{P} - p}{\sqrt{p(1-p)/n}}.$$

tiene distribución aproximada $N(0,1)$.

NOTAS.

1. El **error estándar** de $\overline{P}$ es: $\sigma_{\overline{P}} = \sqrt{\frac{p(1-p)}{n}}$

2. **Si la población es finita de tamaño N** y el muestreo es sin reposición el error estándar (desviación estándar de la hipergeométrica) es:

$$\sigma_{\overline{P}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}.$$

Observe que si el tamaño N de la población, es grande con respecto a n, el factor de corrección $\frac{N-n}{N-1}$ se aproxima a la unidad.

3. Si n es suficientemente grande ($n \geq 30$)

$$P\left[\overline{P} \leq c\right] \cong P\left[Z \leq \frac{c-p}{\sigma_{\overline{P}}}\right].$$

Sin embargo aproximaciones satisfactorias se obtienen si se introduce el **factor de corrección por continuidad** $\frac{1}{2n}$. Luego,

$$P\left[\overline{P} \leq c\right] \cong P\left[Z \leq \frac{(c+1/(2n))-p}{\sigma_{\overline{P}}}\right].$$

4. Observe que la variable estandarizada Z de la proporción es la misma de la binomial X donde, $\overline{P} = X/n$.

$$Z = \frac{X-np}{\sqrt{np(1-p)}} = \frac{\overline{P}-p}{\sqrt{p(1-p)}}.$$

Z tiene distribución aproximada $N(0,1)$, si $n \geq 30$

EJEMPLO 8.6.

Una gran corporación estima en 4% el porcentaje de clientes cuyos créditos han pasado a cobranza dudosa. Un auditor revisa periódicamente las cuentas por cobrar para efectivizar la cobranza.

a) Calcule aproximadamente la probabilidad de que en una muestra aleatoria de 150 clientes con cuentas a cobrar, el 6% haya pasado a cobranza dudosa

b) El auditor acepta que es 4% el porcentaje de clientes cuyos créditos han pasado a cobranza dudosa, si cada vez que revisa una muestra de 100 cuentas por cobrar encuentra menos de 5 cuentas que han pasado a cobranza dudosa. ¿Cuál es la probabilidad de que el auditor acepte que es 4% el porcentaje de clientes cuyos créditos han pasado a cobranza dudosa si realmente el porcentaje es 6%?

SOLUCION

a) Sea $\overline{P} = X/150$, la proporción de clientes cuyos créditos han pasado a cobranza dudosa, en la muestra de 150. La variable X, es el número de clientes cuyos créditos han pasado a cobranza dudosa en la muestra de 150. X es $B(150, 0.04)$. Si se utiliza el modelo binomial para ejecutar el cálculo, se tiene:

$$P[\overline{P}=0.06]=P\left[\frac{X}{150}=\frac{9}{150}\right]=P[X=9]=C_9^{150}(0.04)^9(0.96)^{141}.=0.0688$$

El cálculo, aproximando a la normal, por el teorema central del límite, se puede realizar de dos modos:

Uno, es utilizar la aproximación binomial a la normal. En este caso,

$$Z=\frac{X-np}{\sqrt{np(1-p)}}\frac{X-150\times 0.04}{\sqrt{150\times 0.04\times 0.96}}=\frac{X-6}{2.4} \text{ es aproximadamente } N(0,1)$$

Luego, $P[\overline{P}=0.06]=P[X=9]\cong P[8.5\leq X\leq 9.5]$

$$\cong P\left[\frac{8.5-6}{2.4}\leq Z\leq\frac{9.5-6}{2.4}\right]=P[1.04\leq Z\leq 1.46]=0.0771$$

Otro modo, es usar la distribución muestral de la proporción, $\overline{P}=X/150$, que también se aproxima a la normal. En este caso,

$$Z=\frac{\overline{P}-p}{\sqrt{p(1-p)/n}}=\frac{\overline{P}-0.04}{\sqrt{0.04\times 0.96/150}}=\frac{\overline{P}-0.04}{0.016} \text{ es aproximadamente } N(0,1)$$

$$P[\overline{P}=0.06]=P\left[0.06-\frac{1}{2(150)}\leq\overline{P}\leq 0.06+\frac{1}{2(150)}\right]=P[0.0567\leq\overline{P}\leq 0.0633]$$

$$\cong P\left[\frac{0.0567-0.04}{0.016}\leq Z\leq\frac{0.0633-0.04}{0.016}\right]$$

$$\cong P[1.04\leq Z\leq 1.46]=0.0771.$$

b) Sea: $\overline{P}=X/100$, la proporción de clientes cuyos créditos han pasado a cobranza dudosa en la muestra de 100 cuentas por cobrar. La variable X, es el número de clientes cuyos créditos han pasado a cobranza dudosa en la muestra de 100. X es $B(100,0.06)$. Entonces,

$$P[\overline{P}<0.05/p=0.06]=1-P[\overline{P}\geq 0.05/p=0.06]=1-0.6628=0.3372$$

donde, se ha utilizado la aproximación normal de la distribución de $\overline{P}$:

$$P[\overline{P}\geq 0.05/p=0.06]\cong P\left[Z\geq\frac{0.05-0.06}{0.0237}\right]\cong P[Z\geq -0.42]=0.6628$$

EJEMPLO 8.7.

El gerente de ventas de "TVcable" estima en 20% las conexiones domiciliarias clandestinas. ¿Cuál es la probabilidad de que en una muestra de 100 conexiones domiciliarias seleccionada de una población de 1,000 domicilios que tiene "TVcable", más del 30% resulten clandestinas?.

SOLUCION.

Sea X el número conexiones domiciliarias clandestinas en la muestra de 100. Debido a que el muestreo es sin reposición, X tiene distribución de probabilidad hipergeométrica $H(1000, 200, 100)$.

Se debe calcular $P[X>30] = P[31 \leq X \leq 100]$. Este calculo en el modelo exacto puede presentar algunas complicaciones.

Un método alternativo al cálculo en el modelo exacto, es la aproximación de la distribución hipergeométrica a la normal por el teorema del límite central. En este caso:

$$Z = \frac{X - np}{\sqrt{np(1-p)((N-n)/(N-1))}} = \frac{X - 100 \times 0.2}{\sqrt{100 \times 0.2 \times 0.8 \times ((1000-100)/(1000-1))}} = \frac{X - 20}{3.7966}$$

tiene distribución aproximadamente normal $N(0,1)$.

Luego, $\text{Prob}[X>30] = 1 - P[X \leq 30] \cong 0.0043$

en donde, $P[X \leq 30] \cong P\left[Z \leq \frac{30-20}{3.7966}\right] = P[Z \leq 2.63] = 0.9957$

El otro método alternativo es utilizar la distribución muestral de proporciones para poblaciones finitas que también se aproxima a la normal por el teorema central del límite. En este caso, $\overline{P} = X/100$, es la proporción de conexiones domiciliarias clandestinas en la muestra de 100, y

$$Z = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}} = \frac{\overline{P} - 0.2}{\sqrt{\frac{0.2 \times 0.8}{100} \times (\frac{1000-100}{1000-1})}} = \frac{\overline{P} - 0.2}{0.037966}$$

tiene también distribución aproximadamente normal $N(0,1)$.

Luego, $\text{Prob}[X>30] \cong P[\overline{P} > 0.30] = 1 - P[\overline{P} \leq 0.30] = 0.0043$

en donde $P[\overline{P} \leq 0.3] \cong P\left[Z \leq \frac{0.3-0.2}{0.037966}\right] = P[Z \leq 2.63] = 0.9957$

8.2.3 Distribución muestral de $\frac{n\hat{S}^2}{\sigma^2}$

TEOREMA 8.2. (Distribución muestral de la varianza)
Si $X_1, X_2, ..., X_n$ es una muestra aleatoria escogida de una **distribución normal** $N(\mu, \sigma^2)$, y si,

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

es la varianza muestral, entonces,

a) $E(S^2) = \frac{n-1}{n}\sigma^2$

b) $\frac{nS^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$ tiene distribución $\chi^2(n-1)$.

PRUEBA.

a) Probaremos primero que:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

En efecto,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu) + n(\mu - \bar{X})^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\mu - \bar{X})^2 + n(\mu - \bar{X})^2$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Luego, $E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\right) = \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2]$

$$E(S^2)=\frac{1}{n}\left(n\sigma^2\right)-\frac{\sigma^2}{n}=\sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2.$$

b) Probaremos primero que:

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+n\left(\bar{X}-\mu\right)^2$$

En efecto,

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(X_i-\bar{X}+\bar{X}-\mu\right)^2$$

$$=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}\left(X_i-\bar{X}\right)+n\left(\bar{X}-\mu\right)^2$$

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+n\left(\bar{X}-\mu\right)^2, \text{ya que } \ 2(\bar{X}-\mu)\sum_{i=1}^{n}\left(X_i-\bar{X}\right)=0$$

Dividiendo por σ^2 la identidad probada, resulta,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2}+n\frac{(\bar{X}-\mu)^2}{\sigma^2}$$

Por otra parte, dado que $X_i \sim N(\mu,\sigma^2)$, entonces,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$ tiene distribución $\chi^2(n)$ y

$$n\frac{(\bar{X}-\mu)^2}{\sigma^2}=\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$ tiene distribución $\chi^2(1)$

Utilizando métodos avanzados mas allá de este libro, puede demostrarse que estas dos últimas variables son independientes. Luego, por la propiedad reproductiva de la distribución chi-cuadrado resulta que:

$$\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2}$$ tiene distribución $\chi^2(n-1)$

NOTAS.

1) Si definimos la varianza, $\hat{S}^2 = \dfrac{n}{n-1}S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$,

entonces,

$$E(\hat{S}^2) = E\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}E(S^2) = \left(\frac{n}{n-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \sigma^2.$$

2) En $\hat{S}^2 = \dfrac{n}{n-1}S^2$, se tiene que $\dfrac{n}{n-1} \to 1$, cuando $n \to +\infty$,

3) $\dfrac{(n-1)\hat{S}^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$ tiene distribución $\chi^2(n-1)$.

4) Se verifica que: $\hat{S}^2 = \dfrac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X\right)^2}{n(n-1)} = \dfrac{\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2}{n-1}$

EJEMPLO 8.8.

De una población normal con media 100 y varianza $\sigma^2 = 4$ se obtiene una muestra aleatoria de tamaño $n = 15$. Calcule las probabilidades:

a) $P[1.2428 \le S^2 \le 7.7708]$

b) $P[1.3316 \le \hat{S}^2 \le 8.3256]$

SOLUCION.

a) Con $n = 15$ la variable aleatoria $15S^2/\sigma^2$ tiene distribución chi-cuadrado con 14 grados de libertad, entonces,

$$P[1.2428 \le S^2 \le 7.7708] = P[15 \times 0.3107 \le 15S^2/\sigma^2 \le 15 \times 1.9427]$$
$$= P[4.66 \le \chi^2(14) \le 29.14] = 0.99 - 0.01 = 0.98.$$

b) Con $n = 15$ la variable aleatoria $14\hat{S}^2/\sigma^2$ tiene distribución chi-cuadrado con 14 grados de libertad, entonces,

$$P[1.3316 \le \hat{S}^2 \le 8.3256] = P[(14)(0.3329) \le 14\hat{S}^2/\sigma^2 \le (14)(2.0814)]$$
$$= P[4.66 \le \chi^2(14) \le 29.14] = 0.99 - 0.01 = 0.98$$

8.3 OTRAS DISTRIBUCIONES MUESTRALES

8.3.1 Distribución de $\dfrac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}$

TEOREMA 8.3. (Distribución de $\overline{X}$ cuando σ^2 se desconoce)

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida **de una distribución normal** $N(\mu, \sigma^2)$, donde la varianza poblacional σ^2 es desconocida. Entonces, la variable aleatoria:

$$T = \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}.$$

tiene distribución *t*-student con $n-1$ grados de libertad, o $T \sim t(n-1)$

En efecto, se ha verificado que:

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad \text{y} \quad V = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1),$$

además las variables Z y V son independientes.

Entonces, la variable aleatoria:

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}},$$

tiene distribución t-student con $n-1$ grados de libertad.

NOTA. Observe que

$$V = \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1)$$

Entonces, la variable aleatoria:

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n-1}},$$

tiene también distribución t-student con $n-1$ grados de libertad.

EJEMPLO 8.9.

Si $\overline{X}$ es la media y $\hat{S}^2$ es la varianza de una muestra aleatoria de tamaño $n=9$ seleccionada de una población normal con media: $\mu=90$, calcule

$$P[90-1.1183\times\hat{S}\leq\overline{X}\leq 90+1.1183\times\hat{S}]$$

SOLUCION.

La variable aleatoria:

$$T=\frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}=\frac{\overline{X}-90}{\hat{S}/\sqrt{9}}=\frac{3(\overline{X}-90)}{\hat{S}}$$

se distribuye según t-student con 8 grados de libertad, esto es, $T\sim t(8)$. Entonces,

$$P[-1.1183\leq\frac{\overline{X}-90}{\hat{S}}\leq 1.1183]=P\left[-3\times 1.1183\leq\frac{3(\overline{X}-90)}{\hat{S}}\leq 3\times 1.1183\right]$$

$$=P\left[-3.355\leq t(8)\leq 3.355\right]=0.995-0.005=0.99$$

8.3.2 Distribución muestral de la diferencia de dos medias con varianzas poblacionales conocidas

TEOREMA 8.4. Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos **muestras aleatorias independientes** de tamaños n_1 y n_2 seleccionadas respectivamente de las poblaciones 1 y 2 que tienen medias: μ_1 y μ_2 y varianzas (supuestas conocidas): σ_1^2 y σ_2^2. Entonces, la variable aleatoria $\overline{X}_1-\overline{X}_2$ tiene las siguientes propiedades:

a) $\mu_{\overline{X}_1-\overline{X}_2}=E(\overline{X}_1-\overline{X}_2)=E(\overline{X}_1)-E(\overline{X}_2)=\mu_1-\mu_2$.

b) $\sigma^2_{\overline{X}_1-\overline{X}_2}=V(\overline{X}_1-\overline{X}_2)=V(\overline{X}_1)+V(\overline{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$.

c) Para n_1 y n_2 suficientemente grandes, la distribución de la variable aleatoria:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

es aproximadamente normal estándar $N(0,1)$ (Verificar!).

NOTA.

La aproximación de Z a la normal es muy buena si $n_1 \geq 30$ y $n_2 \geq 30$ sin importar si las poblaciones son discretas o continuas y sin importar sus formas.

Pero, si las dos poblaciones son normales, entonces, la media $\overline{X}_1$ es $N(\mu_1, \sigma_1^2/n_1)$ y $\overline{X}_2$ es $N(\mu_2, \sigma_2^2/n_2)$ para $n_1 \geq 2$ y $n_2 \geq 2$ y por la propiedad reproductiva de la normal la distribución de la variable aleatoria $\overline{X}_1 - \overline{X}_2$ es normal con media $= \mu_1 - \mu_2$ y varianza $= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Por tanto, la distribución de la variable aleatoria estandarizada : $Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

es normal estándar $N(0,1)$ para cualquier valor de $n_1 \geq 2$ y $n_2 \geq 2$.

EJEMPLO 8.10.

La empresa agroindustrial "CHIFLE" embolsa chifles de plátano en sus plantas ubicadas en Tarapoto y Piura. Se sabe que los pesos de los productos embolsados en las dos plantas tienen media iguales a 120 gramos y varianzas iguales a 18 gramos². Con frecuencia se realiza el control de los pesos escogiendo muestras aleatorias del mismo tamaño en cada planta. Encuentre el tamaño de la muestra de manera que la diferencia de las dos medias muestrales sea menos de 2 gramos con probabilidad 0.95.

SOLUCION.

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de los pesos de las muestras escogidas en Tarapoto y Piura respectivamente. Los pesos de las poblaciones tienen medias iguales a 120 gramos y varianzas iguales a 18 gramos². Entonces, la diferencia de las dos medias muestrales $\overline{X}_1 - \overline{X}_2$ tiene distribución aproximadamente normal con:

Media: $\mu_1 - \mu_2 = 0$, y varianza: $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{18+18}{n} = \frac{36}{n}$.

Siempre que $n \geq 30$. Luego, la distribución de la variable

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{36/n}}$$ es normal estándar $N(0,1)$.

Se debe hallar el valor de n tal que $P\left[\left|\overline{X}_1 - \overline{X}_2\right| < 2\right] = 0.95$. Entonces,

$$0.95 = P\left[\left|\overline{X}_1 - \overline{X}_2\right| < 2\right] = P\left[|Z| < \frac{2-0}{\sqrt{36/n}}\right] = P\left[-\frac{\sqrt{n}}{3} < Z < \frac{\sqrt{n}}{3}\right],$$

de donde resulta,

$$P\left(Z < \frac{\sqrt{n}}{3}\right) = 0.975, \quad \frac{\sqrt{n}}{3} = 1.96, \ \sqrt{n} = 5.88, \ n = 34.57 \cong 35.$$

8.3.3 Distribución muestral de la diferencia de dos medias con varianzas poblacionales desconocidas

Sea $\overline{X}_1$ la media de una muestra aleatoria de tamaño n_1 extraída de la **población normal** $N(\mu_1, \sigma_1^2)$, y sea $\overline{X}_2$ la media de otra muestra aleatoria de tamaño n_2 extraída de la **población normal** $N(\mu_2, \sigma_2^2)$, **independiente** de la anterior.

A) Varianzas poblacionales iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

En este caso, la variable aleatoria $\overline{X}_1 - \overline{X}_2$ tiene distribución normal

$$N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

y la variable aleatoria estándar:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

tiene distribución normal $N(0,1)$.

Por otra parte,

$$\frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \text{ y } \frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

$$\text{y,} \quad V = \frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2} + \frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2} \sim \chi^2(n_1+n_2-2)$$

Por tanto, la variable aleatoria:

$$T = \frac{Z}{\sqrt{\dfrac{V}{n_1+n_2-2}}} = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\hat{S}_C^2}{n_1}+\dfrac{\hat{S}_C^2}{n_2}}},$$

tiene distribución *t*-student con n_1+n_2-2 grados de libertad. Donde, **la varianza común**, $\hat{S}_C^2$ tiene la expresión:

$$\hat{S}_C^2 = \frac{(n_1-1)\hat{S}_1^2+(n_2-1)\hat{S}_2^2}{n_1+n_2-2}$$

B) Varianzas poblacionales diferentes $\sigma_1^2 \neq \sigma_2^2$

En este caso la variable aleatoria:

$$T = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\hat{S}_1^2}{n_1}+\dfrac{\hat{S}_2^2}{n_2}}},$$

tiene distribución *t*-student con *g* grados de libertad, donde,

$$g = \frac{\left[\dfrac{\hat{S}_1^2}{n_1}+\dfrac{\hat{S}_2^2}{n_2}\right]^2}{\dfrac{\left[\dfrac{\hat{S}_1^2}{n_1}\right]^2}{n_1-1}+\dfrac{\left[\dfrac{\hat{S}_2^2}{n_2}\right]^2}{n_2-1}}.$$

Dado que *g* rara vez es un entero, se redondea al entero más cercano.

8.3.4 Distribución muestral de la diferencia de dos proporciones

Sean $X_1, X_2, ... X_{n_1}$ e $Y_1, Y_2, ..., Y_{n_2}$ dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones independientes de Bernoullí $B(1, p_1)$ y $B(1, p_2)$, donde p_1 y p_2 son las proporciones poblaciones de éxito respectivos.

Sean las proporciones muestrales de éxito

$$\overline{P}_1 = \frac{X}{n_1} \quad \text{y} \quad \overline{P}_2 = \frac{Y}{n_2}$$

donde las variables aleatorias $X = \sum_{i=1}^{n} X_i$, $Y = \sum_{i=1}^{n} Y_i$ tienen distribuciones binomiales respectivas. $X \sim B(n_1, p_1)$ e $Y \sim B(n_2, p_2)$.

Entonces, la variable aleatoria diferencia de proporciones muestrales: $\overline{P}_1 - \overline{P}_2$ tiene una distribución de probabilidad cuyas propiedades son las siguientes:

a) Media: $\mu_{\overline{P}_1 - \overline{P}_2} = E(\overline{P}_1 - \overline{P}_2) = E(\overline{P}_1) - E(\overline{P}_2) = p_1 - p_2$.

b) Varianza: $\sigma^2_{\overline{P}_1 - \overline{P}_2} = V(\overline{P}_1 - \overline{P}_2) = V(\overline{P}_1) + V(\overline{P}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$.

c) **Para** n_1 y n_2 **suficientemente grandes**, la variable aleatoria estandarizada:

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sigma_{\overline{P}_1 - \overline{P}_2}},$$

tiene distribución de probabilidades aproximadamente normal $N(0,1)$, donde el ***error estándar*** $\sigma_{\overline{P}_1 - \overline{P}_2}$ está dado por la expresión :

$$\sigma_{\overline{P}_1 - \overline{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

EJEMPLO 8.11.

Dos amigos A y B juegan "cara" o "sello" con una moneda. Suponga que en este juego, cada uno lanza la moneda 35 veces y que uno de ellos gana si obtiene 7 caras más que el otro. Calcule la probabilidad de que B gane el juego.

SOLUCION.

Sea X el número de caras que saca el jugador A en las 35 tiradas y sea Y el número de caras que saca el jugador B en las 35 tiradas, entonces, cada variable tiene distribución $B(35,0.5)$, donde $p = 0.5$ es la probabilidad de obtener cara en cada tirada de los jugadores.

Sean $\overline{P}_1 = X/35$ y $\overline{P}_2 = Y/35$ las proporciones de caras de los jugadores A y B respectivamente. El jugador B gana el juego si saca 7 caras más que A o si su proporción de caras es 7/35 = 0.2 más que A. Luego, la probabilidad de que B gane el juego es:

$$P[\overline{P}_2 - \overline{P}_1 > 0.2] = 1 - P[\overline{P}_2 - \overline{P}_1 \leq 0.2] = 1 - P[Z \leq 1.67] = 0.0475,$$

donde: $$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (0.5 - 0.5)}{\sqrt{\frac{(0.5)(0.5)}{35} + \frac{(05)(0.5)}{35}}} = \frac{(\overline{P}_1 - \overline{P}_2) - 0}{0.11952}$$

8.3.5 Distribución muestral de la razón de dos varianzas: $\hat{S}_1^2/\hat{S}_2^2$

TEOREMA 8.5. Si $\hat{S}_1^2$ y $\hat{S}_2^2$ son las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas de dos **poblaciones normales** $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ respectivas, entonces, la variable aleatoria

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución F con grados de libertad $n_1 - 1$ y $n_2 - 1$, esto es, $F \sim F(n_1 - 1, n_2 - 1)$.

En efecto, la variable aleatoria $U = (n_1 - 1)\hat{S}_1^2/\sigma_1^2$ se distribuye como Chi^2 con $n_1 - 1$ grados de libertad y la variable aleatoria $V = (n_2 - 1)\hat{S}_2^2/\sigma_2^2$ se distribuye como Chi^2 con $n_2 - 1$ grados de libertad. Además son independientes. Luego la variable:

$$F = \frac{U/(n_1-1)}{V/(n_2-1)} = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2},$$

se distribuye según F con grados de libertad $n_1 - 1$ y $n_2 - 1$.

Observe que si $\sigma_1^2 = \sigma_2^2$, entonces, $F = \hat{S}_1^2 / \hat{S}_2^2$ se distribuye según F con grados de libertad $n_1 - 1$ y $n_2 - 1$, o $F \sim F(n_1 - 1, n_2 - 1)$

EJEMPLO 8.12.

De poblaciones $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$ independientes se extraen dos muestras de tamaños 13 y 21 respectivamente, hallar el valor de k tal que:

a) $P[\hat{S}_1^2 < k\hat{S}_2^2] = 0.99$

b) $P[\hat{S}_1^2 / \hat{S}_2^2 < k] = 0.05$

SOLUCION.

a) $P[\hat{S}_1^2 < k\hat{S}_2^2] = 0.99$, implica $P[F < k] = 0.99$, donde $F = \hat{S}_1^2 / \hat{S}_2^2 \sim F(12, 20)$.
Entonces, $k = 3.23$.

b) $P[\hat{S}_1^2 / \hat{S}_2^2 < k] = 0.05$, implica $P[F < k] = 0.05$,
donde $F = \hat{S}_1^2 / \hat{S}_2^2 \sim F(12, 20)$. Entonces, $k = 1/f_{0.95,20,12} = 1/2.54 = 0.394$.

EJEMPLO 8.13

Sea: X_1, X_2, una muestra aleatoria de tamaño 2 escogida de una población normal $N(0,1)$,

a) Determine la distribución de: $F = \dfrac{(X_1^2 + X_2^2)/2}{(X_1 - X_2)^2}$

b) Calcule la probabilidad $P[F > 800]$

SOLUCION

La variable aleatoria: $X_1^2 + X_2^2$, es una: $\chi^2(2)$.

La variable aleatoria: $X_1 - X_2$ es $N(0, 2)$, y $(X_1 - X_2)^2/2$, es una : $\chi^2(1)$

a) Por lo tanto, $F = \dfrac{(X_1^2 + X_2^2)/2}{(X_1 - X_2)^2}$ tiene distribución $F(2, 1)$

b) $P[F > 800] = 0.025$

EJERCICIOS

Una media

1. Un taller tiene 5 empleados. Los salarios diarios en dólares de cada uno de ellos son: 5, 7, 8, 10, 10.
   a) Determine la media y la varianza de la población
   b) Halle la distribución muestral de las medias para muestras de tamaño 2 escogidas (sin sustitución) de esta población
   c) Determine la media y la varianza de la distribución muestral de las medias de tamaño 2
   d) Compare la media de las medias muestras con la media de la población. También compare la dispersión de las medias de las muestras con la dispersión de la población.

Rp. a) μ=8, σ^2=3.6, b) 10 muestras, 10 medias, c) $\mu_{\overline{X}} = 8$, $\sigma^2_{\overline{X}} = 10.8/8$, c)

2. De la historia sacada de los registros de la Universidad se ha determinado que las calificaciones del curso de MATE1 y de FILO1 se distribuyen normalmente con las medias respectivas 12 y 15 y con varianzas homogéneas igual a 4. ¿Cuál es la probabilidad de que la media de las notas de un alumno que llevó tales cursos esté entre 13 y 16?

Rp. 0.5984

3. Si $\overline{X}$ denota la media de la muestra aleatoria $X_1, X_2, \ldots, X_9$ de tamaño 9 escogida de la población (X) normal $N(6,6^2)$,
   a) Describa la distribución de probabilidades de la variable aleatoria $\overline{X}$
   b) Halle el percentil 80 de la distribución de $\overline{X}$.
   c) Si $Y = 3X - 5$, calcular $P[\overline{Y} > 28]$.

Rp. a) $N(6, 36/9)$, b) $P_{80} = 7.68$, c) 0.0062

4. Una compañía agroindustrial ha logrado establecer el siguiente modelo de probabilidad discreta de los sueldos (X) en cientos de dólares de su personal:

x	1	2	3	4	5
$f(x) = P[X = x]$	0.1	0.2	0.4	0.2	0.1

si de esta población de sueldos se toman 30 sueldos al azar,
   a) Halle la media y la varianza de la media muestral.
   b) Calcule la probabilidad de que la media muestral esté entre 260 y 330 dólares.

Rp. a) 3, 0.04, b) 0.9104

5. La demanda diaria de un producto puede ser: 0, 1, 2, 3, 4 con probabilidades respectivas: 0.3, 0.3, 0.2, 0.1, 0.1.
   a) Describa el modelo de probabilidad de la demanda promedio de 36 días.
   b) ¿Qué probabilidad hay de que la demanda promedio de 36 días esté entre 1 y 2 inclusive?.

Rp. a) Aproximadamente normal $N(1.4, 1.64/36)$, b) 0.9668.

6. Una empresa comercializadora de café sabe que el consumo mensual (en Kgr) de café por casa está normalmente distribuida con una media desconocida μ y una desviación estándar de 0.30. Si se toma una muestra aleatoria de 36 casas y se registra su consumo de café durante un mes, ¿cuál es la probabilidad de que la media de la muestra esté entre los valores $\mu - 0.1$ y $\mu + 0.1$?.

Rp. 0.9544.

7. La distribución de las notas del examen final de Mat.I resultó ser normal $N(\mu, \sigma^2)$, con cuartiles 1 y 3 iguales a 6.99 y 11.01 respectivamente.
   a) Determine la media y la varianza de la distribución de las notas
   b) Halle el intervalo $[a, b]$ centrado en μ tal que $P[a \leq \overline{X} \leq b] = 0.9544$. donde $\overline{X}$ es la media de la muestra X_1, X_2, X_3, X_4 escogida de esa población.

Rp. a) $\mu = 9$, $\sigma = 3$, b) $a = 6$, $b = 12$.

8. La vida útil (en miles de horas) de una batería es una variable aleatoria X con función de densidad:

$$f(x) = \begin{cases} 2 - 2x & 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

Si $\overline{X}_{36}$ es la media de la muestra aleatoria $X_1, X_2, \ldots, X_{36}$ escogida de X. ¿con qué probabilidad $\overline{X}_{36}$ es mayor que 420 horas?

Rp. 0.0136.

9. Sea $\overline{X}_{40}$ la media de la muestra aleatoria $X_1, X_2, \ldots, X_{40}$ de tamaño $n = 40$ escogida de una población X cuya distribución es geométrica con función de probabilidad:

$$f(x) = \frac{1}{5}\left(\frac{4}{5}\right)^{x-1}, x = 1, 2, \ldots$$

Halle la probabilidad de que la media muestral difiera de la media poblacional en a lo más el 10% del valor de la varianza de la población

Rp. 0.9954.

10. El tiempo de vida de una batería es una variable aleatoria X con distribución exponencial de parámetro: $1/\theta$. Se escoge una muestra de n baterías
   a) Halle el error estándar de la media muestral $\bar{X}$
   b) Si la muestra aleatoria es de tamaño $n = 64$, ¿con qué probabilidad diferirá $\bar{X}$ del verdadero valor de θ en menos de un error estándar?
   c) ¿Qué tamaño de muestra mínimo sería necesario para que la media muestral $\bar{X}$ tenga un error estándar menor a un 5% del valor real de θ.?
   d) Asumiendo muestra grande, ¿qué tamaño de muestra sería necesario para que $\bar{X}$ difiera de θ en menos del 10% de θ con 95% de probabilidad?

Rp. a) $\theta/(n)^{1/2}$, b) $P[\,|\hat{\theta}-\theta|<\theta/(n)^{1/2}]=P[\,|Z|<]=0.6826$, c) $\theta/(n)^{1/2}<0.05\theta$, $n>400$.
d) $P[\,|\hat{\theta}-\theta|<0.1\theta]=0.95$, $P[\,|Z|<0.1(n)^{1/2}]<0.95$, $0.1(n)^{1/2}=1.96$, $n=385$.

11. La utilidad por la venta de cierto artículo, en miles de soles, es una variable aleatoria con distribución normal. En el 5% de las ventas la utilidad ha sido menos que 3.42, mientras que el 1% de las ventas ha sido mayor que 19.32. Si se realizan 16 operaciones de ventas, ¿cuál es la probabilidad de que el promedio de la utilidad por cada operación esté entre \$10,000 y \$12,000?.

Rp. $\mu = 10$, $\sigma = 4$, 0.4772.

12. La vida útil de cierta marca de llantas radiales es una variable aleatoria X cuya distribución es normal con $\mu = 38{,}000$ Km. y $\sigma = 3{,}000$ Km.
   a) Si la utilidad Y (en \$) que produce cada llanta está dada por la relación: $Y = 0.2X + 100$, ¿cuál es la probabilidad de que la utilidad sea mayor que 8,900\$?
   b) Determine el número de tales llantas que debe adquirir una empresa de transporte para conseguir una utilidad media de al menos 7541\$ con probabilidad 0.996.

Rp. a) 0.0228, b) $n = 100$

13. Un proceso automático llena bolsas de café cuyo peso neto tiene una media de 250 gramos y una desviación estándar de 3 gramos. Para controlar el proceso, cada hora se pesan 36 bolsas escogidas al azar; si el peso neto medio está entre 249 y 251 gramos se continúa con el proceso aceptando que el peso neto medio es 250 gramos y en caso contrario, se detiene el proceso para reajustar la máquina.
   a) ¿Cuál es la probabilidad de detener el proceso cuando el peso neto medio realmente es 250?.
   b) ¿Cuál es la probabilidad de aceptar que el peso neto promedio es 250 cuando realmente es de 248 gramos?.

Rp. a) 0.0456, b) 0.0228

14. La utilidad por la venta de cierto artículo, en miles de soles, es una variable aleatoria con distribución normal. En el 5% de las ventas la utilidad ha sido menos que 6,71 mientras que el 1% de las ventas ha sido mayor que 14.66. Si se realizan 16 operaciones de ventas, ¿cuál es la probabilidad de que el promedio de la utilidad por cada operación esté entre \$10,000 y \$11,000?.

Rp. a) X=utilidad $X \sim N(\mu,\sigma^2)$, $0.05=P[X<6.71]$, $0.01=P[X>14.16]$, entonces $\mu=9.793$, $\sigma=1.8742$, $P[10 \leq \overline{X} \leq 11] = P[0.44 \leq Z \leq 2.58]=0.3251$.

15. La vida útil en meses de una batería es una variable aleatoria X con distribución exponencial de parámetro β tal que, $P[X>5/X>2] = e^{-0.6}$
    a) Halle el valor de β y la función de densidad
    b) ¿Cuántas baterías serán necesarias para que duren al menos 20 años con probabilidad 0.0228?

Rp. a) $\beta=0.2$, d) $(240-5n)/(5n)^{1/2}=2$, $n=36$

16. En cierta población de matrimonios el peso en kilogramos de esposos y esposas se distribuye normalmente $N(80,100)$ y $N(64,69)$ respectivamente y son independientes. Si se eligen 25 matrimonios al azar de esta población, calcular la probabilidad de que la media de los pesos sea a lo más 137 Kg.

Rp. 0.0036.

17. Una empresa vende bloques de mármol cuyo peso se distribuye normalmente con una media de 200 kilogramos.
    a) Calcule la varianza del peso de los bloques, si la probabilidad de que el peso esté entre 165 Kg. y 235 Kg. es 0.9876.
    b) ¿Qué tan grande debe ser la muestra para que haya una probabilidad de 0.9938 de que el peso medio de la muestra sea inferior a 205 Kg.?

Rp. a) $\sigma=14$, b) $n=49$.

18. La duración en horas de una marca de tarjeta electrónica se distribuye exponencialmente con un promedio de 1000 horas.
    a) Halle el tamaño n de la muestra de manera que sea 0.9544 la probabilidad de que su media muestral esté entre 800 y 1200 horas.
    b) Si se obtiene una muestra aleatoria de 100 de esas tarjetas calcular la probabilidad que la duración media de la muestra sea superior a 1,100 horas.

Rp. a) n = 100, b) 0.1587.

19. Un procesador de alimentos envasa café en frascos de 400 gramos. Para controlar el proceso se seleccionan 64 frascos cada hora. Si su peso medio es inferior a un valor critico K, se detiene el proceso y se registra. En caso contrario, se continúa la operación sin detener el proceso. Determinar el valor de K de modo que haya una probabilidad de sólo el 5% de detener el proceso

cuando está envasando a un promedio de 407.5 gramos con una desviación estándar de 2.5 gramos.

Rp. ES= 2.5/8=0.3125, $(K-407.5)/0.3125=-1.645$, $K=406.986$.

20. Un proceso para llenar cerveza en botellas de 620 ml. sufre una pérdida en el contenido que tiene una media de 5 ml. y una desviación estándar de 1.2 ml.. Se escogen al azar 36 de tales botellas. Si la meda de la muestra está entre 4.5 y 5.5ml. se acepta que $\mu=5$ ml., en caso contrario; se rechaza que $\mu=5$. ¿Cuál es la probabilidad de aceptar que $\mu=5$ cuando realmente es $\mu=4.8$ ml.?.

Rp. 0.9330.

21. Una empresa de cerveza tiene tres fábricas, una en Cusco, otra en Arequipa y una en Alemania. El 40% de la cerveza es producida en la fábrica de Cusco, el 50% en la de Arequipa y el resto en la de Alemania. La cantidad de cerveza en cada botella es una variable aleatoria X, que para la fábrica de Cusco se distribuye normal con media 620 mililitros y una desviación estándar de 3 mililitros, para la fábrica de Arequipa se distribuye normal con media 623 mililitros y una desviación estándar de 5 mililitros, mientras que para la fábrica de Alemania se distribuye normal con media 621 mililitros y una desviación estándar de 3.5 mililitros.
Se rechaza una cerveza que tenga menos de 615 mililitros o más de 630 mililitros.

a) ¿Con qué probabilidad es rechazada una botella de cerveza?.

b) Calcule la probabilidad de que en 50 cajas de cerveza, menos de 40 botellas sean rechazadas.

Rp. a) $p=0.4\times0.0479+0.5\times0.1356+0.1\times0.0287=0.09183$, b) $X\sim B(600, p)$, $P[X\leq39]=P[Z\leq(39.5-55.098)/7.073779]=P[Z\leq-2.21`]=0.0136$.

22. Una empresa comercializa fardos de algodón cuyo peso X se distribuye normalmente con una media de 250 Kg. y una desviación estándar de 4 Kg. El costo por fardo está dado por: $Y=aX+52$. Halle el valor de a, si se sabe que la media de los costos de 4 fardos escogidos al azar es siempre mayor que \$3100 con probabilidad 0.0228

Rp. $a=12$.

23. Definimos la variable aleatoria "error muestral", por: $|\overline{X}-\mu|$. De todas las muestras de tamaño 36 escogidas al azar de la población $N(\mu,324)$,

a) ¿Qué porcentaje tendrán un error muestral mayor de 4.5?.

b) ¿Para qué valor de k el 95% tienen error muestral no mayor que k?.

Rp. a) 0.1336, b) $k=5.88$.

24. El costo de producción en dólares de un objeto es 100 veces el valor numérico de su longitud. Suponga que la longitud en metros del objeto es una variable aleatoria con distribución normal $N(0.012, 1.44\times10^{-4})$.

    a) ¿Cuál es la distribución del costo medio por objeto si se toman al azar n ($n \geq 2$) objetos?.

    b) Si el precio de venta de cada objeto es \$2.00, calcular la probabilidad de que la utilidad promedio por objeto de 36 objetos tomados al azar, sea a lo más \$0.5.

Rp. a) $N(1.2, 1.44/n)$, b) $\overline{U}_{36} \cong N(0.8, 0.04)$, $P[\overline{U}_{36} \leq 0.5]=0.0668$.

25. Un analista de investigación de mercado toma una muestra aleatoria de 36 clientes de una tienda, de un conjunto de 400 clientes que adquirieron un cupón especial. El monto de las compras mensuales de los 400 clientes constituye una población finita con una media de 2,500 dólares y una desviación estándar de \$660. ¿Cuál es la probabilidad de que la media de la muestra supere los \$2765?.

Rp. población finita, 0.0059

26. Un auditor quiere tomar una muestra aleatoria de una población que consiste de 10,000 cuentas por cobrar, donde $\sigma = \$2000$. ¿De que tamaño debe escoger la muestra si se quiere tener una probabilidad del 95% de que la diferencia entre la media muestral y la media poblacional no exceda el valor \$192?

Rp. Población finita, $n \cong 400$

27. La calificación en una prueba de aptitud es una variable aleatoria X que tiene distribución normal con media igual a 100.

    a) Si se supone que la desviación estándar de todas las calificaciones es $\sigma = 15$, ¿cuántas calificaciones se deben escoger para que la media muestral esté en el intervalo de 90.2 a 109.8 con probabilidad 0.95?.

    b) Si se escogen al azar 16 calificaciones y se encuentra que la desviación estándar $\hat{s} = 12$, ¿cuál es la probabilidad de que la media muestral se encuentre entre 92.194 y 104.023?.

Rp. a) $n = 9$, b) 0.89.

28. Un fabricante afirma que las baterías que produce duran en promedio tres años. En el control de calidad se verifican 16 baterías y si el valor de t calculado: $t_c = (\bar{x} - 3)/(\hat{s}/\sqrt{n})$ está entre $-t_{0.05}$ y $t_{0.05}$, el fabricante está satisfecho con su afirmación. ¿Qué conclusión sacará el fabricante si la muestra da una media de 3.8 años y una desviación estándar $\hat{s} = 1.5$ años?. Suponga que la duración de las baterías tiene distribución normal.

Rp. $t_c = 2.13 \notin [-1.753, 1.753]$, es un producto mejor de lo afirmado.

Una Proporción.

29. Suponga que el 40% de los votos de los electores de una ciudad favorecen al candidato A.
   a) Si se selecciona una muestra aleatoria de 600 electores de la ciudad, ¿cuál es la probabilidad de que la proporción muestral de votos a favor de A esté entre 37% y 45%?
   b) ¿Qué tamaño de muestra se debería escoger si se quiere tener una probabilidad igual a 0.97 de que la proporción de votos a favor de A en la muestra no se diferencie de la proporción supuesta en más del 2%?.

Rp. a) $P[-1.5 \leq Z \leq 2.5]=0.9270$, b) $n=282534=2826$.

30. Una empresa que hace estudios de mercado quiere obtener una muestra aleatoria suficientemente grande de manera que la probabilidad de que la proporción obtenida a favor de un cierto producto resulte inferior al 35% sea igual a 0.0062.
   a) Calcule el tamaño de la muestra a tomar si se supone que la verdadera proporción a favor del producto es $p = 0.4$.
   b) Con el tamaño de muestra calculado en a) y si se supone verdadero el valor del parámetro $p = 0.2$, determinar el intervalo $[a,b]$ centrado en p tal que $\overline{P} \in [a,b]$ con probabilidad 0.95

Rp. a) $n = 600$, b) [0.1608, 0.2392]

31. Un fabricante afirma que a lo más el 2% de todas las piezas producidas son defectuosas. Al parecer esta información es exagerada, por lo que se selecciona una muestra aleatoria de 400 de tales piezas. Si la proporción muestral de defectuosos es mayor que 3% se rechaza la afirmación, en caso contrario se acepta la afirmación.
   a) ¿Cuál es la probabilidad de rechazar la afirmación cuando realmente el 2% de todas las piezas producidas son defectuosas?
   b) ¿Cuál es la probabilidad de aceptar la afirmación cuando realmente el 4% de todas las piezas producidas son defectuosas?

Rp. a) 0.0764, b) 0.1539.

32. El director de la bolsa de trabajo de la universidad afirma que el 60% de los egresados consigue empleo con una remuneración mayor a los $500. Para comprobar esta afirmación se escoge una muestra aleatoria de 600 egresados de la universidad. Si 330 o más pero no más de 390 de la muestra consiguen trabajo con remuneración mayor a los $500, se aceptará la afirmación. En caso contrario se rechazará tal afirmación.
   a) ¿Cuál es la probabilidad de rechazar la afirmación cuando ésta es realmente verdadera.?.

b) ¿Cuál es la probabilidad de aceptar la afirmación cuando realmente el 70% de todos los egresados consiguen trabajo con remuneración mayor a los \$500?.

Rp. a) 0.0124, b) 0.0038.

33. Para controlar la calidad en un proceso de producción de cierto bien de consumo, se seleccionan al azar 46 unidades del bien cada día. Si la proporción de objetos defectuosos en la muestra es al menos $\overline{p}_0$, se detiene el proceso, de otro modo se continua con el proceso. Determine aproximadamente el valor de $\overline{p}_0$ para que con probabilidad de 0.9332 no se continúe con el proceso, cuando la producción total contenga 8% de objetos defectuosos. Rp. 0.02.

34. Un nuevo producto va a salir al mercado si por lo menos el $\overline{p}_0$ (100%) de n personas encuestadas, aceptan el producto. Calcular los valores de n y $\overline{p}_0$ de manera que haya una probabilidad de 0.1112 de que el producto no saldrá al mercado cuando realmente el 58% lo aceptan y una probabilidad de 0.0228 de que el producto saldrá al mercado cuando realmente el 50% lo aceptan.

Rp. $n \cong 400$, $\overline{p}_0 \cong 0.55$.

35. Por experiencia el departamento de créditos de una tienda comercial sabe que sus ventas se pagan con: dinero en efectivo, con cheque o al crédito, con probabilidades respectivas; 0.3, 0.3, y 0.4.. La probabilidad de que una venta sea por más de \$50 es igual a 0.2 si ésta es en efectivo, es igual a 0.9 si ésta es con cheque y es igual a 0.6 si ésta es al crédito.
Si se escoge una muestra aleatoria de 256 personas que ingresan a la tienda, ¿cuál es la probabilidad de que el porcentaje de personas que hayan comprado por más de \$50 sea al menos 50%?. Rp. 0.9881.

36. Un transbordador transporta 300 pasajeros. Se sabe que el peso de la población de pasajeros tiene una media de 63 kgs. y una varianza de 135 kgs² y que el 30% de toda la población de pasajeros tiene pesos que superan los 90 kilogramos.

a) Si el reglamento de seguridad establece que el peso total de los pasajeros del transbordador no debe exceder los 19000 kgs en más del 5% de las veces. ¿Cumple el transbordador las reglamentaciones de seguridad?.

b) Si la población consiste de 2000 pasajeros, ¿qué probabilidad hay de que menos del 25% de los 300 pasajeros tengan un peso que supere los 90 kilogramos?

Rp. a) $Y_{300} \sim N(18900, (201.246)^2)$, $P[Y_{300}>19000]=P[Z>(19000-18900)/201.246]=P[Z>0.50]=0.309$ No cumple, b) $X \sim H(2000, 600,300)$, $P[X\leq 74]=P[Z\leq(74.5-90)/7.3196]=P[Z\leq -2.12]=0.0170$.

37. Un lote grande está formado por dos tipos de artículos: A y B. Los pesos de los artículos de tipo A tienen distribución $N(21, 1)$ y los de tipo B, tienen distribución $N(20,1)$
   a) Si del lote se escoge un artículo al azar, ¿cuál es la probabilidad de que pese entre 19 y 21 si el 80% son de tipo A y el resto son de tipo B?
   b) Si se eligen 200 artículos del lote, ¿cuál es la probabilidad aproximada de que al menos 115 de ellos pesen entre 19 y 21?

Rp. X: Peso, a) $E=\{19\leq X\leq 21\}$, $p=P(E)=P(A)P(E/A)+P(B)P(E/B)=$ $0.8\times 0.4772+0.2\times 0.6826=0.51828$,
b) $Y\sim B(200$, $P[Y>115]=P[Z>(115-103.656)/7.06634]=P[Z>1.61]=0.0537$.

38. Se sabe que el tiempo (en minutos) que lleva realizar una prueba de ecografía en el servicio médico de la PUCP es una variable X que tiene distribución normal con media 30.
   a) Halle la desviación estándar de la distribución si en el 2.28% de los casos las ecografías duran más de 40 minutos.
   b) ¿Qué probabilidad hay de que el tiempo medio de 16 ecografías no sea mayor a 28 minutos?.
   c) Si la proporción de ecografías que indican quiste es 0.10, ¿qué probabilidad hay de que de 10 ecografías a lo más uno indique quiste?

Rp. a) $Z=(40-30)/\sigma=2$, da $\sigma=5$, b) $P[\text{Media}<28]=P[Z<-1.6]=0.0548$,
c) $X\sim B(10, 0.1)$, $P[X\leq 1]=0.7361$

39. De 3000 empleados de una empresa se escoge una muestra aleatoria de 300 empleados para una encuesta sobre condiciones laborales. ¿Cuál es la probabilidad de que la proporción muestral a favor de las condiciones laborales esté comprendido en el intervalo 0.76 y 0.84, si se estima en 80% del total de empleados el porcentaje a favor de las condiciones laborales?.

Rp. 0.9328

40. Una empresa encuestadora debe seleccionar una muestra aleatoria de una población que consiste de 3000 electores para una encuesta de opinión. La empresa estima en 30% del total, el porcentaje a favor de cierto candidato. ¿De que tamaño debe escoger la muestra si se quiere tener una probabilidad del 95% de que la diferencia de la proporción a favor del candidato en la muestra y en la población no exceda el valor 0.0492?

Rp. $n\cong 300$

Varianzas

41. Si $X_1, X_2,..., X_8$ son ocho variables aleatorias independientes y distribuidas cada una normal $N(10,32)$, calcular la probabilidad de que la varianza muestral $S^2=\Sigma(X_i-\overline{X})^2/8$ sea menor o igual que 56.28.

Rp. 0.95.

42. Calcular la probabilidad de que una muestra aleatoria de tamaño 13 escogida de una población normal con varianza $\sigma^2 = 4$ tenga una varianza muestral $\hat{S}^2$, a) menor que 7.01, b) entre 1.19 y 2.1.

Rp. a) 0.95, b) 0.09.

43. Si $X_1, X_2, ..., X_9$ son 9 variables aleatorias independientes y con distribución normal $N(8, 4)$, calcule la probabilidad $P[1.09 \le \hat{S}^2 \le 10.045,\ 7 \le \overline{X} \le 9]$ ($\overline{X}$ y $\hat{S}^2$ son independientes).

Rp. 0.8361.

44. Utilizando la tabla de la distribución F hallar:
a) $F_{0.95,10,15}$, b) $F_{0.99,15,9}$, c) $F_{0.05,30,8}$, d) $F_{0.01,15,9}$

Rp. a) 2.54, b) 4.96, c) 0.4405, d) 0.257.

45. Dos muestras aleatorias independientes de tamaños 21 y 9 respectivamente se toman de una misma población que está normalmente distribuida, ¿cuál es la probabilidad de que la varianza de la primera muestra sea al menos el cuádruple de la varianza de la segunda?.

Rp. 0.025

46. Dos muestras aleatorias independientes de tamaños 7 y 13 respectivamente se toman de una misma población que está normalmente distribuida, ¿cuál es la probabilidad de que la varianza de la primera muestra sea mayor o igual al triple de la varianza de la segunda muestra?.

Rp. 0.05.

47. Sean. $X_1 \sim \chi^2(9)$, $X_2 \sim \chi^2(20)$ y $X = (X_1/9)/(X_2/20)$ hallar los valores a y b tales que:

$$P[a \le X \le b] = 0.925 \text{ y } P[X \le a] = 0.05.$$

Rp. b=2.84, a=1/2.94=0.34

48. Sea $X_1, ..., X_{10}$ una muestra aleatoria escogida de una población normal $N(0,1)$,

a) Halle la distribución de $F = (\sum_{i=1}^{10} X_i^2)/10 \Big/ (\sum_{i=1}^{5} X_i^2)/5$,

b) Calcule la probabilidad $P[F < 1/3.33]$

Rp. a) $F \sim F(10,5)$, b) 0.05

49. Sea X_1, X_2 una muestra aleatoria escogida de una población normal $N(0,1)$,

a) Determine la distribución de: $F = \left[\dfrac{X_1 + X_2}{X_1 - X_2}\right]^2$

b) Calcule la probabilidad $P[F < 161]$

Rp. a) $F \sim F(1,1)$, b) 0.95

Diferencia de dos medias

50. Para comparar la duración promedio (en meses) μ_1 y μ_2 de dos marcas de baterías B1 y B2 se escogen dos muestras aleatorias independientes de tamaños respectivos $n_1 = 32$ y $n_2 = 36$. Si la media muestral de B1 es mayor que la media muestral de B2 en mas de 2 meses, se acepta que $\mu_1 > \mu_2$. En caso contrario se acepta que $\mu_1 = \mu_2$. Calcular la probabilidad de aceptar que $\mu_1 > \mu_2$ cuando realmente $\mu_1 = \mu_2$. Suponga que las varianzas de las duraciones de B1 y B2 son respectivamente $\sigma_1^2 = 16$ y $\sigma_2^2 = 9$.

Rp. 0.0104

51. Una firma comercializadora afirma que el peso promedio (en gramos) μ_1 y μ_2 de dos marcas de café instantáneo C1 y C2, es el mismo. Para verificar la afirmación se escogen dos muestras aleatorias independientes de tamaños 36 sobres de cada marca.. Si la media muestral de C1 es mayor que la media muestral de C2 en mas de 0.5 gramos, se rechaza que $\mu_1 = \mu_2$. En caso contrario, se acepta que $\mu_1 = \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 = \mu_2$ cuando realmente $\mu_1 = \mu_2 + 2$?. Suponga que las varianzas de las poblaciones C1 y C2 son respectivamente $\sigma_1^2 = 9$ y $\sigma_2^2 = 4$

Rp. 0.0062

52. El jefe de compras está por decidir si comprar una marca A o una marca B de focos para la compañía. Para ayudarle a optar por una de ellas se escogen dos muestras aleatorias de tamaños $n_1 = 10$ y $n_2 = 9$ focos respectivamente de las marcas A y B, resultando, las desviaciones estándares respectivas $\hat{s}_1 = 200$ y $\hat{s}_2 = 150$. Si la diferencia entre las medias muestrales es mayor que 173 horas, se acepta que $\mu_1 \neq \mu_2$. En caso contrario, se acepta que $\mu_1 = \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 \neq \mu_2$ cuando realmente $\mu_1 = \mu_2$?. Se asume

que la vida útil de ambas marcas tiene distribución normal con varianzas iguales.

Rp. $T \sim t(17)$, 0.05.

53. Para comparar los salarios que se pagan a los empleados en dos grandes empresas E1 y E2 se escogen dos muestras aleatorias independientes de tamaños $n_1 = 16$ y $n_2 = 13$ respectivamente de E1 y E2 resultando las desviaciones estándares respectivas $\hat{s}_1 = \$120$ y $\hat{s}_2 = \$55$. Si la diferencia entre las medias muestrales no es mayor que 65\$, se acepta que $\mu_1 = \mu_2$. En caso contrario, se acepta que $\mu_1 \neq \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 \neq \mu_2$ cuando realmente $\mu_1 = \mu_2$?. Se asume que los salarios en ambas empresas tienen una distribución normal con varianzas diferentes.

Rp. $T \sim t(22)$, 0.10.

Diferencia de dos proporciones.

54. Dos programas de televisión A y B tienen como ratings (porcentaje de hogares donde se ve el programa) de 40 y 20 respectivamente. Se toma una muestra aleatoria de 300 hogares con T.V. durante la transmisión del programa A y otra de 100 hogares durante la transmisión de B, ¿cuál es la probabilidad de que los resultados muestren que el programa A tiene un rating mayor a la de B en 10%?.

Rp. 0.0207

55. Un fabricante afirma que el 30% de mujeres y el 20% de hombres prefieren su nuevo producto de aseo personal. Si se hace una encuesta a 200 hombres y 200 mujeres elegidos aleatoriamente, ¿con qué probabilidad la proporción muestral de mujeres menos la proporción muestral de hombres está en el intervalo [−19%, 19%]?

Rp. 0.9634.

56. Se escoge una muestra de 600 electores que acaban de votar, entre las 9 a.m. y las 3 p.m. para estimar la proporción de votantes a favor de los candidatos A y B. En una encuesta hecha en la víspera se estimó en 30% y 35% los porcentajes a favor de A y B respectivamente. ¿Cuál es la probabilidad de que la proporción muestral de B exceda a la proporción muestral de A en al menos 10%?.

Rp. 0.0322.

57. Cierta marca de cigarrillos es preferida por el 30% de mujeres y el 25% de hombres. En una encuesta hecha a 300 personas de cada sexo elegidas aleatoriamente, ¿cuál es la probabilidad de que la proporción muestral de mujeres que prefieren esa marca sea mayor a la de los hombres?

Rp. 0.9147

Capítulo 9

ESTIMACION DE PARAMETROS.

9.1 Introducción

Al realizar una investigación estadística a menudo se sabe o se supone que la población (discreta o continua), de la cual se selecciona una muestra aleatoria, tiene una forma funcional específica $f(x)$ cuyo(s) parámetro(s) se intenta determinar. Si el parámetro a determinar es denotado por θ, entonces, la distribución de la población será denotada por $f(x, \theta)$.

Los métodos de inferencia estadística consisten en seleccionar una muestra aleatoria de la población, de manera que a partir de la información que se obtenga de la muestra:

1) Determinar el valor del parámetro desconocido θ, ó

2) Decidir si θ, ó alguna función de θ, es igual a algún valor preconcebido θ_0 de θ.

El primero de estos dos procedimientos se denomina ***estimación del parámetro*** θ. El segundo procedimiento se conoce como ***prueba de hipótesis del parámetro*** θ.

El método de estimación de un parámetro puede ser ***puntual*** o ***por intervalo.*** En el primer caso, la estimación del parámetro θ es un número. Mientras que en el segundo caso la estimación incluye un intervalo en el que están comprendidos los valores del parámetro.

9.2 Estimación puntual de parámetros

Definición. Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño *n* seleccionada de una población cuya distribución es $f(x, \theta)$, siendo θ el parámetro. Se denomina ***estimador puntual*** del parámetro θ a cualquier estadística $\hat{\Theta} = H(X_1, X_2,..., X_n)$, cuyo valor $\hat{\theta} = H(x_1, x_2,..., x_n)$ proporcionará una ***estimación*** del parámetro en cuestión.

Un estimador puntual del parámetro θ es pues, una *variable aleatoria* (función de la muestra) $\hat{\Theta}$, mientras que una estimación puntual es el *valor numérico* $\hat{\theta}$ del estimador.

Por dar un ejemplo, un **estimador** puntual de la media poblacional θ, es la estadística media muestral (variable aleatoria) $\hat{\Theta} = \overline{X}$, cuyo valor numérico $\hat{\theta} = \overline{x}$ es la **estimación** puntual del parámetro θ.

No toda función de la muestra es un *buen estimador* del parámetro, un buen estimador, es aquel que está más cerca del parámetro que se estima. Para que un estimador puntual sea bueno debe tener ciertas propiedades. Una de estas propiedades es que sea *insesgado*, propiedad conocida también como *no-sesgado*, *imparcial*, o *sin vicio*.

9.2.1. Estimador insesgado

Definición. Se dice que la estadística $\hat{\Theta} = H(X_1, X_2,..., X_n)$ es un **estimador insesgado** del parámetro θ si $E(\hat{\Theta}) = \theta$. En caso contrario, se dice que es estimador sesgado.

Si la estadística $\hat{\Theta} = H(X_1, X_2,..., X_n)$ es un estimador insesgado del parámetro θ, entonces, su valor $\theta = H(x_1, x_2,..., x_n)$ es la **estimación insesgada** del parámetro θ.

EJEMPLO 9.1.

Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño *n* extraída de una población cualquiera $f(x, \mu, \sigma^2)$, (discreta o continua). Entonces,

a) La media muestral $\overline{X}$ es un estimador insesgado de la media poblacional μ, ya que

$$E(\overline{X}) = \mu.$$

El valor $\bar{x}$ de $\overline{X}$ es la estimación insesgada de μ.

b) La proporción muestral $\overline{P}$ es un estimador insesgado de la proporción de éxitos p de una población binomial, por que,

$$E(\overline{P}) = p.$$

c) La varianza muestral

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

es un **estimador sesgado** de la varianza poblacional σ^2, ya que

$$E(S^2) = \frac{n-1}{n}\sigma^2.$$

Sin embargo, la estadística,

$$\hat{S}^2 = \frac{nS^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

es un estimador insesgado de la varianza poblacional σ^2, por que,

$$E(\hat{S}^2) = \sigma^2.$$

NOTA. El cálculo la varianza por $\hat{S}^2$ es denominado **método de muestra** mientras que el cálculo de la varianza por S^2 es denominado **método de población**. Observar que cuando n es suficientemente grande el coeficiente $(n-1)/n$ tiende a uno.

EJEMPLO 9.2.

a) La diferencia $\overline{X}_1 - \overline{X}_2$ de dos medias muestrales es estimador insesgado de la diferencia $\mu_1 - \mu_2$ de dos medias poblaciones. Ya que,

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2.$$

b) La diferencia de dos proporciones muestrales $\overline{P}_1 - \overline{P}_2$ es un estimador insesgado de la diferencia de dos proporciones de éxitos binomiales $p_1 - p_2$. Puesto que,

$$E(\overline{P}_1 - \overline{P}_2) = p_1 - p_2.$$

9.2.2. Estimador eficiente

Definición. Si hay dos o más estimadores puntuales insesgados de un parámetro θ, se denomina estimador **más eficiente** a aquel estimador que tenga *menor varianza*.

EJEMPLO 9.3.

Sea X_1, X_2, X_3, X_4 una muestra aleatoria de cualquier población con distribución $f(x, \mu, \sigma^2)$. Dados los estimadores del parámetro μ:

$$\text{a) } \hat{\Theta}_1 = \frac{X_1 + X_2 + X_3 + X_4}{4}, \quad \text{y} \quad \text{b) } \hat{\Theta}_2 = \frac{4X_1 - X_3 + X_4}{4}$$

verificar que la media de la muestra; es el estimador más eficiente de μ.

SOLUCION.

Ambos estimadores son insesgados. En efecto,

$$E(\hat{\Theta}_1) = \frac{4\mu}{4} = \mu \quad \text{y} \quad E(\hat{\Theta}_2) = \frac{4\mu - \mu + \mu}{4} = \frac{4\mu}{4} = \mu$$

Por otra parte, las varianzas respectivas son:

$$V(\hat{\Theta}_1) = \frac{4\sigma^2}{16} = \frac{\sigma^2}{4} \quad \text{y} \quad E(\hat{\Theta}_2) = \frac{16\sigma^2 + \sigma^2 + \sigma^2}{16} = \frac{18\sigma^2}{16}$$

Luego, la estadística $\hat{\Theta}_1$ que está definida como la media de la muestra; es el estimador más eficiente de μ.

Uno de los métodos para determinar estimadores puntuales es el de *máxima verosimilitud* que se describe a continuación.

9.2.3. Método de máxima verosimilitud

Supongamos que una población X está distribuida como $f(x,\theta)$, en donde θ es el parámetro que tratamos de estimar.

El procedimiento para determinar el estimador de máxima verosimilitud es como sigue:

1) Elegir una muestra aleatoria $X_1, X_2, ..., X_n$ de la población y determinar la distribución conjunta de la muestra en sus valores observados respectivos $x_1, x_2, ..., x_n$. Esta función del parámetro θ conocida también como **función de verosimilitud** está dada por:

$$L(\theta) = f(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) ... f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

2) El valor de θ que maximiza a la función $L(\theta)$, es la **estimación** de máxima verosimilitud (EMV) de θ. Este valor denotaremos por

$$\hat{\theta} = H(x_1, x_2, ..., x_n).$$

La estadística correspondiente $\hat{\Theta} = H(X_1, X_2, ..., X_n)$ es el **estimador** de máxima verosimilitud de θ.

3) A menudo se usa $L = \ln(L(\theta))$. En este caso el valor de θ que maximiza a $L(\theta)$ es la solución $\hat{\theta}$ de la ecuación:

$$\frac{dL}{d\theta} = 0.$$

4) Si la distribución de probabilidad de la población contiene k parámetros $\theta_1, \theta_2, ..., \theta_k$, la función de verosimilitud está dada por:

$$L(\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2, ..., \theta_k).$$

La estimación de máxima verosimilitud de cada parámetro θ_i es la solución $\hat{\theta}_i$ de la ecuación respectiva:

$$\frac{\partial L}{\partial \theta_1} = 0, \ \frac{\partial L}{\partial \theta_2} = 0, ..., \ \frac{\partial L}{\partial \theta_k} = 0$$

donde $L = \ln(L(\theta_1, \theta_2, ..., \theta_k))$.

EJEMPLO 9.4.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n, seleccionada de una población con distribución binomial $B(1,p)$. Hallar el estimador del parámetro p usando el método de máxima verosimilitud.

SOLUCION.

La distribución de probabilidad de cada variable aleatoria X_i es:

$$f(x_i, p) = p^{x_i}(1-p)^{1-x_i}, \ x_i = 0, 1.$$

La función de verosimilitud de la muestra aleatoria es entonces,

$$L(p)=\prod_{i=1}^{n} f(x_i,p)=(p)^{\Sigma x_i}(1-p)^{n-\Sigma x_i} \qquad (*)$$

Luego, $$L=\ln(L(p))=(\ln(p))\sum_{i=1}^{n} x_i+(\ln(1-p))\left(n-\sum_{i=1}^{n} x_i\right).$$

Derivando la función L con respecto a p e igualando a cero, da:

$$\frac{dL}{dp}=\frac{\sum_{i=1}^{n} x_i}{p}-\frac{n-\sum_{i=1}^{n} x_i}{1-p}=0.$$

De donde resulta que:

$$\hat{p}=\frac{\sum_{i=1}^{n} x_i}{n}=\overline{p}.$$

Luego, la proporción muestral $\overline{p}$ es la estimación ($\overline{P}$ es el estimador) de máxima verosimilitud de la proporción poblacional p, siempre que:

$$\sum_{i=1}^{n} x_i \neq 0, \quad \text{y} \quad \sum_{i=1}^{n} x_i \neq n.$$

Observar que:

si $\sum_{i=1}^{n} x_i=0$, entonces, de (*) resulta que el E.M.V. de p es $\hat{p}=0$.

si $\sum_{i=1}^{n} x_i=n$, entonces, de (*) resulta que el E.M.V. de p es $\hat{p}=1$.

EJEMPLO 9.5.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria escogida de una población con distribución $N(\mu,\sigma^2)$. Hallar el estimador de la media μ y la varianza σ^2 de la distribución, usando el método de máxima verosimilitud.

SOLUCION.

La distribución de probabilidades de la población normal de parámetros μ y σ^2 asociada a cada variable aleatoria X_i está dada por:

$$f(x_i,\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

La función de verosimilitud es:

$$L(\mu,\sigma^2)=\left[f(x_i,\mu,\sigma^2)\right]^n=\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right]^n.$$

$$L(\mu,\sigma^2)=\left(2\pi\sigma^2\right)^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

Luego,

$$L=\ln\left(L(\mu,\sigma^2)\right)=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Derivando la función L con respecto a μ e igualando a cero da:

$$\frac{dL}{d\mu}=-0+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

de donde resulta: $$\hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}=\overline{x}.$$

Luego, la media muestral $\overline{x}$ es la estimación ($\overline{X}$ es el estimador) de máxima verosimilitud de la media poblacional μ.

Por otro lado, derivando la función L con respecto a σ^2 e igualando a cero da:

$$\frac{dL}{d\sigma^2}=-\frac{1}{2}\frac{n}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

de donde resulta: $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = s^2.$$

Luego, la varianza muestral s^2 es la estimación (S^2 es el estimador) de máxima verosimilitud de σ^2.

NOTA. Los estimadores de máxima verosimilitud tienen la ***propiedad de invarianza***, en el sentido de que **si $\hat{\theta}$ es E.M.V del parámetro θ, entonces, cualquier expresión $g(\hat{\theta})$ es E.M.V. de $g(\theta)$**. En el ejemplo 9.5, la varianza muestral s^2 es la estimación de máxima verosimilitud de la varianza poblacional σ^2, luego, la desviación estándar muestral s es la estimación de máxima verosimilitud de la desviación estándar σ.

EJEMPLO 9.6.

Para estimar la vida media de un tipo de componente electrónica se selecciona una muestra aleatoria de 10 unidades, se les somete a prueba y se encuentra que 6 de ellas siguen funcionando después de 3,000 horas. Suponiendo que la vida útil de las componentes es una variable aleatoria T con distribución exponencial de parámetro β, estimar la vida media de tales componentes producidas.

SOLUCION.

Sea p la proporción poblacional de todas las componentes que siguen funcionando después de 3,000 horas. Entonces,

$$p = P[T > 3000] = \int_{3,000}^{+\infty} \beta e^{-\beta t} dt = e^{-3000\beta}$$

De donde resulta: $$\hat{p} = e^{-3000\hat{\beta}}.$$

Por otra parte, la estimación por máxima verosimilitud de la proporción p es (ejemplo 9.4):

$$\hat{p} = \overline{p} = 6/10.$$

Por tanto, utilizando la propiedad de invarianza en $\hat{p} = e^{-3000\hat{\beta}}$, resulta:

$$\hat{\beta} = -\frac{1}{3,000}\ln(\hat{p}) = -\frac{1}{3,000}\ln(\frac{6}{10}) \quad \text{y} \quad \hat{\bar{x}} = \frac{1}{\hat{\beta}} = 5,872.85 \text{ horas}$$

EJERCICIOS

1. De una población $f(x, \mu, \sigma^2)$ se escogen dos muestras aleatorias independientes de tamaños n_1 y n_2. Sean $\overline{X}_1$ y $\overline{X}_2$, y $\hat{S}_1^2$ y $\hat{S}_2^2$ sus medias y varianzas respectivas.

   a) Si $\overline{X} = (n_1\overline{X}_1 + n_2\overline{X}_2)/(n_1 + n_2)$, ¿es la estadística $\overline{X}$ un estimador insesgado del parámetro μ?.

   b) Si $\hat{S}_c^2 = \dfrac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$, ¿es la estadística $\hat{S}_c^2$ un estimador insesgado del parámetro σ^2?.

   Rp. a) si, b) si.

2. Si $\overline{P}_1$ y $\overline{P}_2$ son las proporciones de dos muestras de tamaños n_1 y n_2 escogidas de una población Bernoullí $B(1, p)$, verifique que la estadística $\hat{P} = \dfrac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}$ es un estimador insesgado del parámetro p.

3. Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras independientes de tamaños n_1 y n_2 respectivamente escogidas de una población X de Poisson con parámetro λ.

   a) Probar que la estadística $\hat{\theta} = \dfrac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$ es un estimador insesgado de del parámetro λ.

   b) Hallar la varianza del estimador.

   Rp. a) $E(\hat{\theta}) = \lambda$, b) $V(\hat{\theta}) = \dfrac{\lambda}{n + n_2}$..

4. La duración, en horas, de cierta clase de foco sigue una distribución exponencial con media desconocida θ horas. Se toma una muestra de un sólo foco al azar y se mide su duración X en horas. Si con X se estima θ, ¿se podría decir que X es un estimador insesgado de θ?.

   Rp. Si por que $X \sim \text{Exp}(1/\theta)$, entonces, $E(X) = (1/\theta)^{-1} = \theta$.

5. Dos métodos diferentes e independientes dieron lugar a dos estimadores insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$ del parámetro θ. Las desviaciones estándares de estos estimadores son 0.4 y 0.6 respectivamente. Los estimadores son combinados de la siguiente manera:

$$\hat{\theta} = r\hat{\theta}_1 + (1-r)\hat{\theta}_2 \qquad 0 < r < 1.$$

Hallar el valor de r que haga mínima la varianza del estimador $\hat{\theta}$.

Rp. r=0.6923.

6. Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño n de una población de Bernoullí $B(1, p)$. De las siguientes estadísticas:

$$\hat{\Theta}_1 = \frac{\sum_{i=1}^{n} X_i - X_k}{n-1}, \qquad \hat{\Theta}_2 = \frac{\sum_{i=1}^{n} X_i^2}{n}$$

a) ¿Cuáles son estimadores insesgados del parámetro p ?.
b) ¿Cuál de ellas es de varianza mínima?.

Rp. a) ambos, b) 2do.

7. Sea $X_1, X_2,..., X_{50}$ una muestra aleatoria de tamaño 50 escogida de una población con distribución geométrica de parámetro p, $0 < p < 1$,

$$P[X = x] = p(1-p)^x, \quad x = 0,1,2,...$$

a) Determinar el estimador de máxima verosimilitud para p.
b) Estimar p, si $x_1 + x_2 + \cdots + x_{50} = 100$.

Rp. a) $1/(\bar{x}+1)$, b) $1/3$.

8. Sea $X_1, X_2,..., X_{20}$ una muestra aleatoria de tamaño 20 escogida de una población con distribución binomial de parámetro p, $0 < p < 1$,

$$P[X = x] = C_x^2 p^x (1-p)^{2-x}, \quad x = 0, 1, 2$$

Determinar el estimador de máxima verosimilitud para p, si en la muestra el valor 0 ocurre 4 veces, el valor 1 ocurre 9 veces y el valor 2 ocurre 7 veces

Rp. $L(p) = 2^9 p^{23} (1-p)^{17}$, $\hat{p} = 23/40$.

9. El número de ventas diarias de cierta mercadería es una variable aleatoria X de Poisson con un promedio de λ ventas por día.
a) Si $X_1, X_2,..., X_{50}$, son las ventas de 50 días, estimar λ, por el método de máxima verosimilitud.
b) Si en los 50 días se han hecho 30 ventas de tal mercadería, estimar el promedio λ de ventas diarias.

Rp. a) $\hat{\lambda} = \bar{x}$, b) 30/50.

10. De una población de variable aleatoria continua X se extrae una muestra aleatoria $X_1, X_2, \ldots, X_n$, y se define la variable aleatoria Bernoullí:

$$Y = \begin{cases} 1 & \text{si } X_i > 0 \\ 0 & \text{si } X_i \leq 0 \end{cases}$$

a) Usando el método de máxima verosimilitud estimar la proporción p de todos los valores positivos, esto es, estimar, $p = P[X > 0]$.

b) Estime el valor de p si una muestra aleatoria de tamaño 80 de X ha dado 64 valores positivos y 16 valores negativos.

c) Si $X \sim N(\mu, 0.04)$, utilizando a) y b), calcular aproximadamente el valor de μ.

Rp. a) $\hat{p} = \sum y_i / n$, b) 0.8, c) $\mu = 0.168$.

11. El tiempo, en meses, que dura una componente electrónica es una variable aleatoria T de distribución exponencial con parámetro β. Para estimar β se prueban 30 componentes y se encuentra que 18 fallan antes de los 6 meses.

a) Utilizando el método de máxima verosimilitud, estimar la proporción de todas las componentes que fallan antes de los 6 meses.

b) Utilice el resultado de a) para estimar por máxima verosimilitud de β.

Rp. a) $\hat{p} = 18/30$, b) $\hat{p} = 1 - e^{-6\hat{\beta}}$, $\hat{\beta} = -\ln(12/30)/6$.

12. La longitud de cierto tipo de objeto producidos por una máquina, puede estar por arriba o por abajo de la medida estándar de 2 pulgadas. Suponga que tal longitud tiene distribución normal $N(\mu, 0.0025)$.

a) Utilizando el método de máxima verosimilitud estime la proporción p de todos los objetos cuya longitud está por arriba de 2 pulgadas.

b) Si en una muestra de 1,000 de tales objetos se encontró que 992 tenían longitud por arriba de 2 pulgadas, utilizando a) estime la media de la longitud de todos los objetos producidos.

Rp. a) $\hat{p} = \overline{p}$, b) $\overline{p} = 0.992 = P[Z > (2 - \hat{\mu})/0.05]$, $\hat{\mu} = 2.1205$.

13. Una máquina produce objetos cuyo peso en gramos tiene distribución normal $N(30, \sigma^2)$, con σ desconocido. Los objetos son defectuosos si el peso es menor que 26 o mayor que 34 gramos. Para estimar σ se pesa un objeto cada vez hasta que un defectuoso sea obtenido. Hallar el estimador de máxima verosimilitud de σ si en un control el primer defectuoso se halló en la décima prueba.

Rp. $1 - \hat{p} = P[-4/\hat{\sigma} \leq Z \leq 4/\hat{\sigma}]$, $\hat{p} = 0.1$, $\hat{\sigma} = 2.43$.

9.3 Estimación de parámetros por intervalos

9.3.1 Intervalo de confianza

Una estimación de punto no nos dice cuán próximo está la estimación al parámetro que se estima, por lo tanto, no es muy significativa, sin no se tiene alguna medida del error que se comete en la estimación. Es deseable pues tener cierto *grado de confianza* de que la estimación de punto se halle dentro de cierta variación.

La *estimación por intervalo* (propuesto por J. Neyman en 1937), es la estimación de un parámetro θ dentro de un intervalo de extremos cerrados $[a,b]$, donde los números a y b se obtienen a partir de la distribución de la estadística que estima puntualmente el parámetro; y a partir de los valores de la muestra.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n escogida de una población $f(x,\theta)$, cuyos valores experimentales (o datos) respectivos son $x_1, x_2, ..., x_n$. Sea además, la variable aleatoria $\hat{\Theta} = H(X_1, X_2, ..., X_n)$ una estadística para estimar el parámetro θ cuya distribución de probabilidad sea conocida. Si dado el número $1-\alpha$, y si a partir de la distribución de $\hat{\Theta}$ se pueden encontrar las variables aleatorias $A = H_1(X_1, X_2, ..., X_n)$, y $B = H_2(X_1, X_2, ..., X_n)$ tales que:

$$P[A \leq \theta \leq B] = 1-\alpha$$

entonces, se dice que *el intervalo aleatorio* $[A, B]$ es el ***intervalo estimador*** del parámetro θ con el ***nivel de confianza*** de $(1-\alpha)\times 100\%$, o que $\theta \in [A, B]$ con probabilidad $1-\alpha$.

Además, si $a = H_1(x_1, x_2, ..., x_n)$, y $b = H_2(x_1, x_2, ..., x_n)$, son los valores numéricos que resultan de sustituir los valores de la muestra en las *estadísticas* A y B respectivamente, entonces, se dice que el *intervalo numérico* $[a,b]$ es *el intervalo de confianza del* $(1-\alpha)\times 100\%$ para θ, o que $\theta \in [a,b]$ con *nivel o grado de confianza del* $(1-\alpha)\times 100\%$.

La diferencia, pues, entre los intervalos $[A, B]$ y $[a,b]$, es que el primero es un intervalo aleatorio y por lo tanto tiene validez afirmar que la probabilidad que contenga al parámetro θ es igual a $1-\alpha$. Mientras que el segundo es un intervalo numérico fijo, y en este caso, *no tiene validez afirmar* que la probabilidad $P[a \leq \theta \leq b] = 1-\alpha$.

La **interpretación** del intervalo de confianza es como sigue: Si a partir de los datos de una muestra aleatoria de tamaño n, hemos construido el intervalo $a \leq \theta \leq b$ con grado de confianza, por ejemplo, del 95% para el parámetro θ, entonces, si se seleccionan repetidamente 100 muestras de tamaño n, tendremos 100 intervalos semejantes al intervalo $[a,b]$, y se confía que 95 de estos 100 intervalos contengan el parámetro θ.

La probabilidad $1-\alpha$, o el porcentaje $(1-\alpha)\times 100\%$ es denominado *el grado (o nivel) de confianza*. Sus valores más utilizados son 0.95, 98, 0.99 entre otros. Al número α se le denomina también *riesgo de estimación* por intervalo.

A los valores a y b se les denomina *límites de confianza* del parámetro θ. El número a es el *límite inferior de confianza* y el número b es el *límite superior de confianza*.

Por otra parte, si la estadística A_1 verifica:

$$P[A_1 \leq \theta] = 1-\alpha$$

se concluye que el intervalo $[a_1, +\infty[$ es un *intervalo de estimación unilateral* del parámetro θ del $(1-\alpha)\times 100\%$, donde a_1 es el valor de A_1 que se obtiene a partir de la muestra.

Similarmente, si la estadística B_1 verifica:

$$P[\theta \leq B_1] = 1-\alpha$$

se concluye que el intervalo $]-\infty, b_1]$ es un *intervalo de estimación unilateral* del parámetro θ del $(1-\alpha)\times 100\%$, donde b_1 es el valor de B_1 que se obtiene a partir de la muestra.

9.4 Intervalo de confianza para la media μ

9.4.1 Intervalo de confianza para la media μ: Varianza σ^2 supuesta conocida

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n seleccionada de una población normal (o de cualquier otro tipo si $n \geq 30$) con media μ y varianza σ^2 supuestamente conocida.

El mejor estimador puntual del parámetro μ es la media muestral $\bar{X}$.

Se utiliza, entonces, la distribución muestral de la media $\overline{X}$ para determinar el intervalo de confianza del parámetro μ.

Si la *población es normal* $N(\mu,\sigma^2)$, entonces, la distribución de la estadística $\overline{X}$ es normal $N(\mu, \sigma^2/n)$ para cualquier valor de n ($n \geq 2$).

Si la *población no es normal*, pero tiene media μ y varianza σ^2 finitas, entonces, siempre que el tamaño n de la muestra sea suficientemente grande, ($n \geq 30$), por el teorema del límite central, la distribución de $\overline{X}$ es *aproximadamente* normal $N(\mu,\sigma^2)$.

Por tanto, según sea el caso, la distribución de la variable aleatoria:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

es *exactamente* (o *aproximadamente*) normal $N(0,1)$.

Luego, dado el valor $1-\alpha$ (o en %), en la distribución de Z, se determinan los valores $\mp z_{1-\alpha/2}$ (figura 9.1) tales que:

$$P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1-\alpha.$$

Sustituyendo $Z = (\overline{X} - \mu)\sigma/\sqrt{n}$, se tiene,

$$P\left[-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1-\alpha.$$

De donde resulta,

$$P\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.$$

$$P[A \leq \mu \leq B] = 1-\alpha.$$

donde: $A = \overline{X} - z_{1-\alpha/2}\,\sigma/\sqrt{n}$ y $B = \overline{X} + z_{1-\alpha/2}\,\sigma/\sqrt{n}$

son variables aleatorias.

Esto es, si $\overline{X}$ es estimador de μ, **se tiene la probabilidad** $1-\alpha$ **de que el *intervalo* (*aleatorio* o *estimador*)** $[A, B]$ contenga al parámetro μ.

Luego,

Si $\bar{x}$ es el valor de la media $\bar{X}$ para una muestra aleatoria de tamaño n escogida de una población con varianza σ^2 supuesta conocida, el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ es:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

El valor $z_{1-\alpha/2}$ se busca en la tabla normal $N(0,1)$, tal que $P[Z \leq z_{1-\alpha/2}] = 1-\alpha/2$.

La ilustración, es la figura 9.1, en la que los valores $a = \bar{x} - z_{1-\alpha/2}\,\sigma/\sqrt{n}$ y $b = \bar{x} + z_{1-\alpha/2}\,\sigma/\sqrt{n}$ son los *límites de confianza* de μ, inferior y superior, respectivamente.

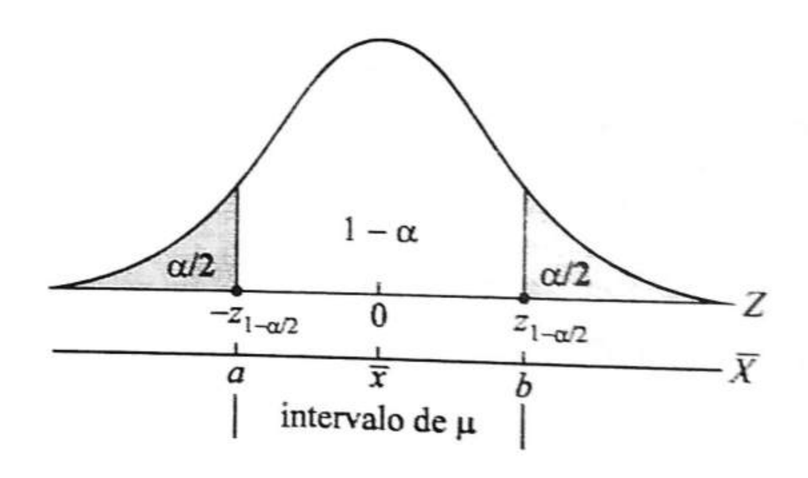


Figura 9.1: Intervalo de estimación para μ (con estadística Z.)

Interpretación. Si se seleccionan repetidamente 100 muestras de tamaño n, y calculamos las medias de cada una de ellas, tendremos 100 intervalos semejantes al intervalo $[a,b]$, y se confía que 95 de estos 100 intervalos contengan el parámetro μ y 5 de los 100 no lo contengan, como se muestra en la figura 9.1b. Los puntos circulares en el centro de cada intervalo indican la estimación puntual del parámetro μ.

Note que todos los intervalos son del mismo ancho, ya que este último sólo depende de $z_{1-\alpha/2}$ una vez que se determina $\bar{x}$.

En la figura 9.1b los intervalos correspondientes a las medias $\bar{x}_2, \bar{x}_4$ no contienen al parámetro μ, mientras que el resto de los intervalos si contienen al parámetro.

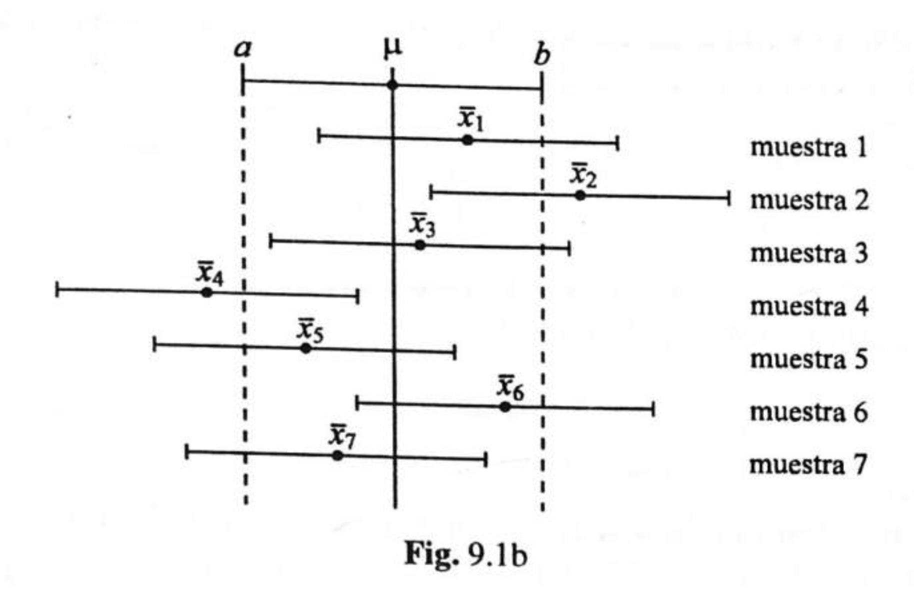


Fig. 9.1b

NOTA. (Población finita, muestreo sin reemplazo)

Si la muestra aleatoria de tamaño n es escogida *sin reposición* de una ***población finita de tamaño N***, entonces, si $n \geq 30$, la variable aleatoria:

$$Z' = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}},$$

tiene distribución aproximadamente normal $N(0,1)$. Utilizando la distribución de Z' se determina el intervalo de confianza de μ.

Luego, si $\overline{x}$ es un valor de la media $\overline{X}$ para una muestra aleatoria de tamaño n escogida de una **población *finita de tamaño N*** con varianza σ^2 supuesta conocida, el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ es:

$$\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \leq \mu \leq \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

NOTA. (Error estándar)

Se denomina **error estándar** de un estimador a la desviación estándar del estimador. A su valor numérico se denomina **error estándar estimado**. Por ejemplo, el **error estándar** (E.S.) de la media de una muestra de una población infinita (o población finita con sustitución) es

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Si la población es finita de tamaño N y el muestreo es sin reposición el **error estándar** (E.S.) de la media muestral es:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Luego, el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ se puede obtener a partir de los **límites de tolerancia o confianza**:

$$\overline{x} \mp z_{1-\alpha/2} E.S.$$

EJEMPLO 9.7.

Una muestra aleatoria de 400 pequeños comerciantes indicó que la media de los ingresos mensuales era de \$800. Estime la media de la población que consiste de los ingresos de todos los pequeños comerciantes mediante un intervalo de confianza del 95%. Asuma que la desviación estándar de esta población es \$200.

SOLUCION.

Sea X la variable que representa el ingreso mensual de todos los pequeños comerciantes, cuya media μ se quiere estimar a partir de una muestra aleatoria de tamaño $n = 400$ escogida de esa población.

La estimación puntual de μ es $\overline{x} = \$800$.

Para el nivel de confianza $1-\alpha = 0.95$, en la tabla normal estándar se encuentra: $z_{1-\alpha/2} = z_{0.975} = 1.96$.

El error estándar de la media $\overline{X}$ es: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{400}} = 10.$

Los límites de confianza de μ son:

$$\overline{x} \mp z_{1-\alpha/2}\sigma_{\overline{X}} = 800 \mp 1.96\times 10 = 800 \mp 19.6$$

Luego, $780.4 \leq \mu \leq 819.6$. Esto es, se tiene una confianza del 95% que la media de los ingresos mensuales de todos los pequeños comerciantes está en el intervalo [\$780.4, \$819.6].

NOTA. Muestras diferentes del mismo tamaño darán diferentes valores de $\overline{x}$, y por tanto darán diferentes intervalos de estimación de μ. Decir que el intervalo de estimación contiene al parámetro con confianza 95%, equivale a decir que 95 por 100 de los intervalos contienen a la media μ y que sólo el 5 por 100 no lo contienen.

EJEMPLO 9.8.

Un analista de investigación de mercados escogió una muestra aleatoria de 100 clientes de una población de 500 clientes de una gran tienda que declaran ingresos mayores a $800. El encontró que los clientes de la muestra gastaban en la tienda un promedio de $1000 por año. Si con este valor de la muestra estimó que el gasto promedio por año de esta población finita varía de $940 a $1060, ¿qué nivel de confianza utilizó?. Suponga que la desviación estándar de tal población es $300.

SOLUCION.

Se tienen los siguientes datos: $\bar{x} = \$1000$, $\sigma = \$300$, $N = 500$, $n = 100$.

El intervalo de confianza del $1-\alpha$ en % para la media μ de esta **población finita** está dado por:

$$\bar{x} - z_{1-\alpha/2}E.S \leq \mu \leq \bar{x} + z_{1-\alpha/2}E.S.$$

donde $E.S$, el error estándar de la media es el número:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{300}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} = 26.8597$$

Dado que: $940 \leq \mu \leq 1060$, entonces, $1060 = 1000 + z_{1-\alpha/2} \times 26.8597$.

Luego, $z_{1-\alpha/2} = 60/26.8597 \cong 2.23$, $\alpha/2 = 0.0129$, y

$$1-\alpha = 0.9742.$$

NOTA. (Error de estimación)

Si μ se estima puntualmente por $\bar{x}$, entonces, el **error de la estimación** es el valor numérico $|\bar{x} - \mu|$ (ver figura que sigue).

error

a $\bar{x}$ μ b

$a = \bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}}$ y $b = \bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}$

El valor mínimo del error de estimación es igual a cero, esto ocurre, cuando $\bar{x}$ estima exactamente a μ.

El valor máximo del error de la estimación es igual a: $z_{1-\alpha/2}\sigma_{\bar{X}}$, ya que del intervalo de estimación de μ resulta:

$$|\bar{x} - \mu| \leq z_{1-\alpha/2}\sigma_{\bar{X}}.$$

Luego,

Si $\overline{x}$ estima a μ, entonces, se tiene una confianza del $(1-\alpha)\times 100\%$ de que el error de la estimación no será superior a $z_{1-\alpha/2}\sigma_{\overline{X}}$, donde $\sigma_{\overline{X}}$ es el error estándar de la media.

Por **ejemplo**, en el ejemplo 9.7 se tiene una confianza del 95% de que al estimar μ por \$800, el error de la estimación no será superior a \$19.6. Esto es, se tiene una confianza del 95%, que la media muestral $\overline{x} = \$800$ difiere de la media verdadera en a lo más \$19.6.

En el ejemplo 9.8, se tiene una confianza del 97.42% de que al estimar μ por \$1000, el error de la estimación no será superior a \$60.

NOTA. (Tamaño de la muestra)

Se puede determinar que tan grande debe ser el tamaño n de la muestra, de manera que si μ se estima por $\overline{x}$, el error de estimación no sea mayor que un valor dado e. En efecto, el valor de n se obtiene de

$$z_{1-\alpha/2}\sigma_{\overline{X}} \leq e$$

Entonces,

Si se utiliza $\overline{x}$ como una estimación de μ, entonces, **se tiene una confianza** del $(1-\alpha)\times 100\%$ **de que el error no será mayor que el valor dado** e cuando el tamaño de la muestra sea al menos

$$n = \frac{(z_{1-\alpha/2}\sigma)^2}{e^2}$$

Si la población es finita de tamaño N y el muestreo es sin sustitución, el error estándar es $\sigma_{\overline{X}} = \left(\sigma/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$ y el valor de n se calcula por:

$$n = \frac{z^2_{1-\alpha/2}\sigma^2 N}{z^2_{1-\alpha/2}\sigma^2 + e^2(N-1)}.$$

Si el valor de n contiene decimales se redondea al número entero más cercano.

Por **ejemplo**:

a) En el ejemplo 9.7, se tiene una confianza del 95% de que al estimar la media de la población, el error de la estimación no será mayor de \$15 cuando el tamaño n de la muestra sea al menos:

$$n = \frac{(z_{1-\alpha/2}\sigma)^2}{e^2} = \frac{(1.96)^2(200)^2}{(15)^2} = 682.95 \cong 683.$$

b) En el ejemplo 9.8, se tiene una confianza del 97% de que al estimar la media de la población, el error de la estimación no será mayor de \$50 cuando el tamaño n de la muestra sea al menos:

$$n = \frac{z_{1-\alpha/2}^2 \sigma^2 N}{z_{1-\alpha/2}^2 \sigma^2 + e^2(N-1)} = \frac{(2.17)^2(300)^2(500)}{(2.17)^2(300)^2 + (50)^2(500-1)} = 126.79 \cong 127.$$

Observe que cuanto menor es el error de la estimación mayor es el tamaño de la muestra requerida.

NOTA. En estadística aplicada es frecuente utilizar el valor $z_{1-\alpha/2} = 2$ (n se obtiene resolviendo: $2\sigma_{\bar{X}} = 2$). El tamaño de la muestra requerida para estimar la media μ de una población finita de tamaño N está dada por la expresión:

$$n = \frac{N\sigma^2}{\sigma^2 + \frac{e^2}{4}(N-1)}$$

Si se desconoce σ^2 pero se conoce el rango de variación de los datos, se puede utilizar: $\sigma \cong \frac{\text{Rango}}{6}$. (También se aproxima por $\sigma \cong$Rango/4)

Si el método de muestreo es por estratos, el tamaño de la muestra requerida para estimar μ con un error máximo de estimación e está dado por:

$$n = \frac{\sum_1^K N_i^2 \sigma_i^2 / w_i}{\frac{e^2}{4} N^2 + \sum_1^K N_i \sigma_i^2}$$

donde w_i es el % de observaciones asignados al estrato i, $i = 1,2,...,k$.

Si se desconoce σ_i^2, pero se conoce el rango de variación de los datos en cada estrato, entonces, $\sigma_i \cong \frac{\text{Rango}_i}{6}$ (Ver ejercicios 17 y 18)

NOTA. (Estimación del total de la población)

Si la muestra aleatoria de tamaño n, se escoge de una población finita de tamaño N, entonces,

$$\textbf{\textit{Total de la población:}} \sum_{i=1}^{N} X_i = N\mu .$$

La estimación puntual del total $N\mu$ es: $N\bar{x}$.

El intervalo de confianza del ($(1-\alpha)100\%$) para μ es:

$$\bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}$$

donde, el error estándar de la media muestral es: $\sigma_{\bar{X}} = \left(\sigma/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$.

Luego, el intervalo de confianza del ($(1-\alpha)100\%$) para $N\mu$ es:

$$N(\bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}}) \leq N\mu \leq N(\bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}).$$

Para dar un **ejemplo**, en el ejemplo 9.8, la estimación puntual del total de gastos de la población $N\mu$ es $N\bar{x} = 500 \times \$1000 = \$500000$

Los límites de confianza al 95% para el total $N\mu$ son:

$$N(\bar{x} \mp z_{1-\alpha/2}\sigma_{\bar{X}}) = 500(\$1000 \mp 1.96 \times 26.8597) = 500000 \mp 26322.506.$$

Luego, $N\mu \in [473{,}677.494,\ 526{,}322.506]$ con confianza 95%.

Consecuentemente, si el total de la población $N\mu$ se estima en \$500,000, se tiene una confianza del 95% de que el error de la estimación no será superior a \$26,322.506.

9.4.2 Intervalo de confianza para la media μ: Varianza σ^2 supuesta desconocida

A) Población no normal

Si la población no es normal pero el tamaño de la muestra es suficientemente grande ($n \geq 30$), se utiliza la desviación estándar $\hat{s}$ de la muestra, como estimación puntual de la desviación estándar σ de la población.

Entonces, utilizando el teorema central del límite, se concluye que el intervalo de confianza del $(1-\alpha) \times 100\%$ para μ es *aproximadamente*:

$$\bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}.$$

donde, el error estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ se sustituye por el error estándar estimado $\hat{\sigma}_{\bar{X}} = \hat{s}/\sqrt{n}$ si el muestreo es con o sin sustitución en una población infinita (con

sustitución en una población finita de tamaño N), y se sustituye por $\hat{\sigma}_{\overline{X}} = \left(\hat{s}/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$ si el muestreo es sin sustitución en una población finita de tamaño N.

B) Población normal

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida de una población normal $N(\mu, \sigma^2)$ donde la varianza σ^2 es supuesta desconocida y sean la media y la varianza muestrales respectivas:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad \hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Se sabe que las variables aleatorias:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}, \qquad V = \frac{(n-1)\hat{S}^2}{\sigma^2},$$

tienen distribuciones respectivas, normal $N(0,1)$ y chi-cuadrado con $n-1$ grados de libertad . Además Z y V son variables aleatorias independientes. Entonces, la variable aleatoria:

$$T = \frac{Z}{\sqrt{\dfrac{V}{n-1}}} = \frac{\overline{X} - \mu}{\dfrac{\hat{S}}{\sqrt{n}}},$$

tiene distribución t-student con $n-1$ grados de libertad, esto es, $T \sim t(n-1)$.

Por tanto, dado el número $1-\alpha$, en la distribución de probabilidad de T se encuentran los números $\pm t_{1-\alpha/2,\,n-1}$ (figura 9.2) tales que:

$$P[-t_{1-\alpha/2,\,n-1} \leq T \leq t_{1-\alpha/2,\,n-1}] = 1-\alpha.$$

Al sustituir la expresión de T se obtiene:

$$P\left[-t_{1-\alpha/2,\,n-1} \leq \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}} \leq t_{1-\alpha/2,\,n-1}\right] = 1-\alpha.$$

$$P[\overline{X} - t_{1-\alpha/2,\,n-1}\,\hat{S}/\sqrt{n} \leq \mu \leq \overline{X} + t_{1-\alpha/2,\,n-1}\,\hat{S}/\sqrt{n}] = 1-\alpha.$$

Luego,

> Si $\bar{x}$ y $\hat{s}$ son la media y la desviación estándar respectivamente para un valor particular $x_1, x_2, ..., x_n$ de la muestra aleatoria de tamaño *n* escogida de la población **normal con varianza σ^2 desconocida**, entonces, el intervalo de confianza de $(1-\alpha)\times 100\%$ para μ es
>
> $$\bar{x} - t_{1-\alpha/2,n-1}\, \hat{s}/\sqrt{n} \le \mu \le \bar{x} + t_{1-\alpha/2,n-1}\, \hat{s}/\sqrt{n}$$

El valor $t_{1-\alpha/2,n-1}$ se encuentra en la tabla t-student con $n-1$ grados de libertad tal que $P[T \le t_{1-\alpha/2,n-1}] = 1-\alpha/2$.

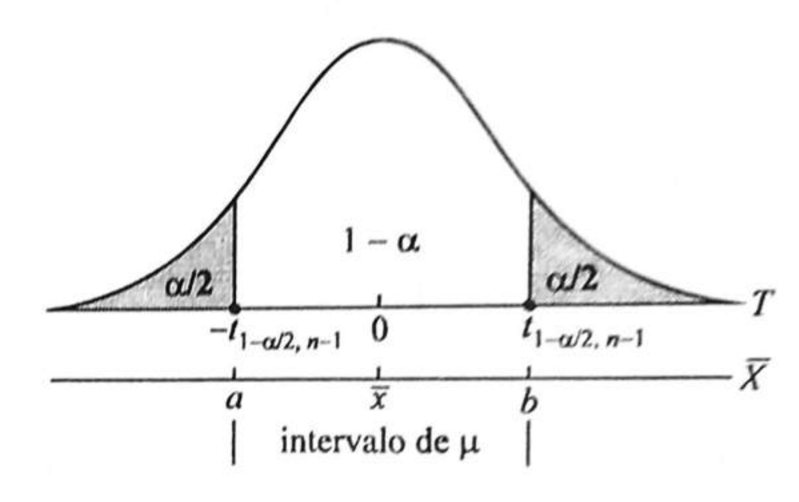


Figura 9.2: Intervalo de estimación para μ (con estadística *t*).

La ilustración es la figura 9.2, donde,

$$a = \bar{x} - t_{1-\alpha/2,n-1}\, \hat{s}/\sqrt{n} \quad \text{y} \quad b = \bar{x} + t_{1-\alpha/2,n-1}\, \hat{s}/\sqrt{n}$$

son los límites de confianza de μ inferior y superior respectivamente.

EJEMPLO 9.9.

Una muestra aleatoria de 10 frascos de conservas de palmito de la empresa agroindustrial "LA PALMA" de Iquitos ha dado los siguientes pesos netos en gramos:

278, 285, 280, 290, 285, 275, 284, 295, 280, 287

a) Estime la media de los pesos netos por frasco de todos los frascos de conserva de palmito de esta empresa utilizando un intervalo de confianza del 95%.

b) ¿Con qué grado de confianza se estima tal promedio en un intervalo cuya longitud es 10.588 gramos?.

Suponga población del peso neto de los frascos de conserva de palmito de esta empresa se distribuye según el modelo de probabilidad normal.

SOLUCION.

a) Sea X el peso neto de los frascos de conserva de palmito, cuya media μ se quiere estimar a partir de esa muestra aleatoria de tamaño $n = 10$.
Se supone que la distribución de X es normal con desviación estándar σ no conocida.
Para el nivel de confianza: $1-\alpha = 0.95$ y grados de libertad: $n-1=9$, en la tabla t-student se encuentra: $t_{1-\alpha/2,n-1} = t_{0.975,9} = 2.262$.
De la muestra se obtiene $\bar{x} = 283.9$ y $\hat{s} = 5.934$.
El error estándar de la media $\bar{X}$ es $\hat{s}/\sqrt{10} = 5.934/3.1623 = 1.8765$.
Los límites de confianza inferior y superior para μ son respectivamente:

$$\bar{x} \mp t_{1-\alpha/2,n-1}\,\hat{s}/\sqrt{n} = 283.9 \mp 2.262 \times 1.8765 = 283.9 \mp 4.2446$$

Luego, $\mu \in [279.66, 288.14]$, con confianza del 95%.

NOTA. El lector debería resolver este problema usando un paquete de computo.

b) Si la longitud del intervalo de estimación es 10.588, entonces, el error máximo de la estimación puntual es 5.294. Con $n = 10$. se debe hallar $1-\alpha$. Entonces,

$$t_{1-\alpha/2,n-1}\,\hat{s}/\sqrt{n} = t_{1-\alpha/2,9} \times 1.8765 = 5.294$$

de donde resulta:

$$t_{1-\alpha/2,9} = 2.821,\ \ 1-\alpha/2 = 0.99,\ \alpha = 0.02 \ \text{y}\ \ 1-\alpha = 0.98.$$

9.5 Intervalo de confianza para la varianza

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n, escogida de una población normal con varianza σ^2, parámetro desconocido.

Un estimador puntual de la varianza σ^2 es la varianza muestral

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

cuyo valor $\hat{s}^2$ es la estimación puntual de σ^2.

Para determinar el intervalo de confianza para la varianza σ^2 se utiliza la estadística:

$$X = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

cuya distribución es chi-cuadrado con $n-1$ grados de libertad, esto es, $X \sim \chi^2(n-1)$ para $n \geq 2$.

Dado el nivel de confianza: $1-\alpha$, en la distribución $\chi^2(n-1)$, se encuentran los valores $\chi^2_{\alpha/2,\,n-1}$ y $\chi^2_{1-\alpha/2,\,n-1}$ tales que (figura 9.3).

$$P[\chi^2_{\alpha/2,\,n-1} \leq X \leq \chi^2_{1-\alpha/2,\,n-1}] = 1-\alpha.$$

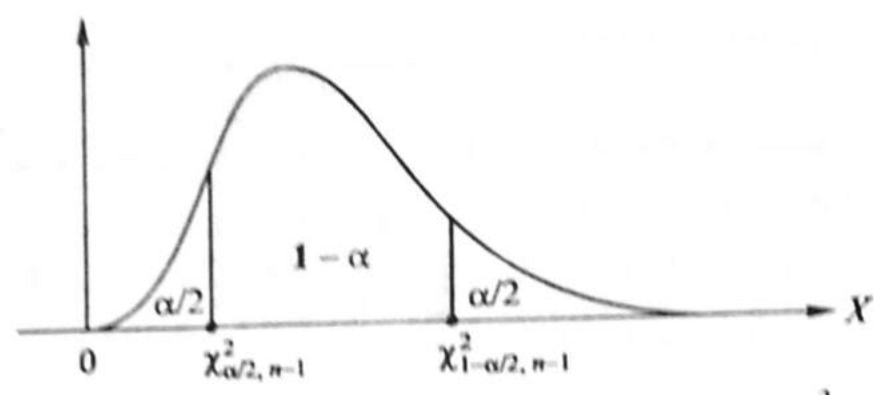


Figura. 9.3 Intervalo de confianza de la varianza σ^2.

Sustituyendo $X = (n-1)\hat{S}^2/\sigma^2$ resulta:

$$P\left[\frac{(n-1)\hat{S}^2}{\chi^2_{1-\alpha/2,\,n-1}} \leq \sigma^2 \leq \frac{(n-1)\hat{S}^2}{\chi^2_{\alpha/2,\,n-1}}\right] = 1-\alpha$$

Luego, si $\hat{s}^2$ es la varianza de una muestra aleatoria de tamaño n seleccionada de una población normal, entonces, el intervalo de confianza de $(1-\alpha)100\%$ para σ^2 es:

$$\frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2,\,n-1}} \leq \sigma^2 \leq \frac{(n-1)\hat{s}^2}{\chi^2_{\alpha/2,\,n-1}}$$

Los valores, $\chi^2_{\alpha/2,\,n-1}$ y $\chi^2_{1-\alpha/2,\,n-1}$ se hallan la tabla chi-cuadrado con $n-1$ grados de libertad y con áreas acumuladas respectivas de $\alpha/2$ y $1-\alpha/2$.

EJEMPLO 9.10.

Una máquina corta automáticamente masas de pan. Para estimar la variabilidad de los pesos de las partes, se tomó una muestra aleatoria de 10 partes cortadas por la máquina, resultando los siguientes pesos en gramos

9.8, 9.9, 10.1, 10.3, 9.9, 10.1, 9.7, 10.3, 10.4, 9.9.

Desarrolle un intervalo de confianza del 95% para la varianza de los pesos de todas las partes cortadas por la máquina.
Suponga que los pesos de todas las partes cortadas se distribuyen según la normal.

SOLUCION.

Con $\alpha = 0.05$, $n = 10$ y $r = n - 1 = 9$ grados de libertad, en la tabla chi-cuadrado se encuentran:

$$\chi^2_{\alpha/2,\, n-1} = \chi^2_{0.025,\, 9} = 2.70 \quad \text{y}$$

$$\chi^2_{1-\alpha/2,\, n-1} = \chi^2_{0.975,\, 9} = 19.02\,.$$

De los datos de la muestra resulta: $\hat{s}^2 = 0.056$.

Los límites de confianza inferior y superior del 95% para la varianza σ^2 son respectivamente:

$$\frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2,\, n-1}} = \frac{9(0.056)}{19.02} = 0.0265,$$

$$\frac{(n-1)\hat{s}^2}{\chi^2_{\alpha/2,\, n-1}} = \frac{(9)(0.056)}{2.70} = 0.1867.$$

Por lo tanto, el intervalo de confianza del 95% para la varianza σ^2 es:

$$0.0265 \le \sigma^2 \le 0.1867.$$

Observe que sacando raíz cuadrada en este intervalo, resulta:

$$0.1628 \le \sigma \le 0.432.$$

que viene a ser **el intervalo de confianza del 95% para la desviación estándar σ de la población de los pesos de todas las partes cortadas.**.

9.6 Intervalo de confianza para la razón de dos varianzas

Sean $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos muestras aleatorias *independientes* de tamaños n_1 y n_2 seleccionadas de dos poblaciones normales respectivas con varianzas σ_1^2 y σ_2^2.

Un estimador puntual de la razón de las varianzas: σ_1^2/σ_2^2 es la estadística $\hat{S}_1^2/\hat{S}_2^2$.

Para determinar el intervalo de confianza de σ_1^2/σ_2^2 se utiliza la estadística F definida por:

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2},$$

que tiene distribución de probabilidad F con grados de libertad $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$. Esto es, $F \sim F(r_1, r_2)$.

En efecto, las variables aleatorias:

$$U = \frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2}, \quad \text{y} \quad V = \frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2}$$

tienen distribuciones respectivas chi-cuadrado con $n_1 - 1$ y $n_2 - 1$ grados de libertad.

Entonces, la variable aleatoria:

$$F = \frac{U/(n_1-1)}{V/(n_2-1)} = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución F con grados de libertad $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$.

Observar que para obtener tal estadística F, no se requiere asumir que las dos poblaciones tengan igual promedio.

Dado el grado de confianza $1-\alpha$, en la distribución $F \sim F(r_1, r_2)$ se pueden encontrar los valores $f_{\alpha/2, r_1, r_2}$ y $f_{1-\alpha/2, r_1, r_2}$ (figura 9.4) tales que:

$$P[f_{\alpha/2, r_1, r_2} \le F \le f_{1-\alpha/2, r_1, r_2}] = 1-\alpha.$$

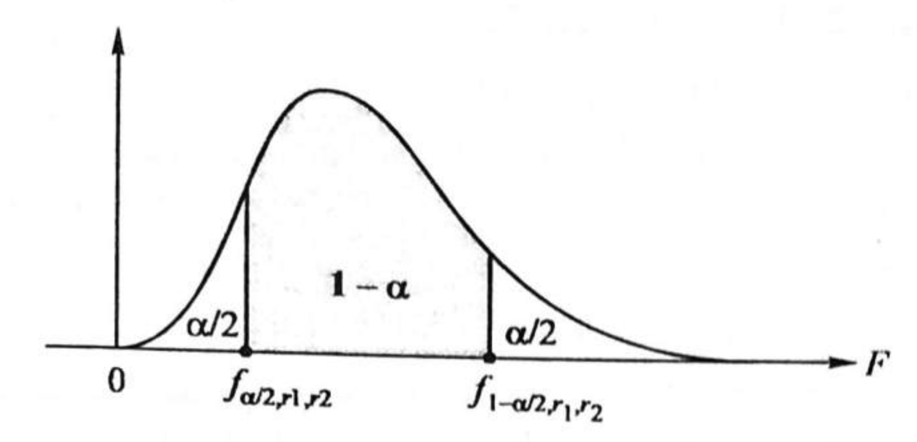


Figura. 9.4. Intervalo de confianza de la varianza σ_1^2/σ_2^2

Sustituyendo $F = (\hat{S}_1^2/\sigma_1^2)/(\hat{S}_2^2/\sigma_2^2)$ y dado que:

$$f_{\alpha/2,\, r_2,\, r_1} = \frac{1}{f_{1-\alpha/2,\, r_1,\, r_2}}$$

$$f_{1-\alpha/2,\, r_2,\, r_1} = \frac{1}{f_{\alpha/2,\, r_1,\, r_2}}$$

resulta,

$$P\left[\frac{\hat{S}_1^2}{\hat{S}_2^2} f_{\alpha/2,\, r_2,\, r_1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\hat{S}_1^2}{\hat{S}_2^2} f_{1-\alpha/2,\, r_2,\, r_1}\right] = 1-\alpha.$$

Luego,

Si $\hat{s}_1^2$ y $\hat{s}_2^2$ son las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones normales, entonces, el intervalo de confianza de $(1-\alpha)\times 100\%$ para σ_1^2/σ_2^2 es:

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} f_{\alpha/2,\, r_2,\, r_1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\hat{s}_1^2}{\hat{s}_2^2} f_{1-\alpha/2,\, r_2,\, r_1}$$

EJEMPLO 9.11.

El gerente de ventas de una cadena de hipermercados quiere comparar la variabilidad de las ventas diarias de dos sucursales A y B. Se sabe que todas las ventas de A y de B se distribuyen normalmente. Dos muestras aleatorias de ventas: Una de 8 días de A y otra de 6 días de B revelaron las siguientes ventas en miles de soles:

Muestra de A: 22, 20, 21, 19, 17, 23, 21, 18.

Muestra de B: 12, 15, 14, 13, 16, 14.

Utilice un intervalo de confianza del 95% para la razón de dos varianzas, para determinar si son iguales o no las varianzas de las dos poblaciones de ventas diarias de A y B?.

SOLUCION.

Sean X, Y las variables aleatorias que representan las ventas de A y de B respectivamente. Se supone que las distribuciones de X, Y son normales.

Con $\alpha = 0.05$, y grados de libertad $r_1 = n_1 - 1 = 7$ y $r_2 = n_2 - 1 = 5$ en la tabla F se encuentran:

$$f_{1-\alpha/2, r_2, r_1} = f_{0.975, 5, 7} = 5.29,$$

$$f_{\alpha/2, r_2, r_1} = 1/f_{1-\alpha/2, r_1, r_2} = 1/f_{0.975, 7, 5} = 1/6.85 = 0.146 .$$

De los datos de la muestra resultan: $\hat{s}_1^2 = 4.125$ y $\hat{s}_2^2 = 2$.

Los límites de confianza del 95% para σ_1^2/σ_2^2 inferior y superior son respectivamente:

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} f_{\alpha/2, r_2, r_1} = \frac{4.125}{2}(0.146) = 0.3011$$

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} f_{1-\alpha/2, r_2, r_1} = \frac{4.125}{2}(5.29) = 10.9106$$

Por lo tanto, el intervalo de confianza del 95% para la varianza σ_1^2/σ_2^2 es:

$$0.3011 \leq \sigma_1^2/\sigma_2^2 \leq 10.9106.$$

Dado que el intervalo contiene a la unidad, es decir, dado que:

$$\sigma_1^2/\sigma_2^2 = 1 \in [0.3011,\ 10.9106] ,$$

se debería inferir con un nivel de confianza del 95%, que las dos varianzas poblacionales son iguales.

9.7 Intervalo de confianza para la diferencia entre dos medias

9.7.1 Intervalo de confianza para la diferencia entre dos medias: Varianzas σ_1^2 y σ_2^2 supuestas conocidas

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos **muestras aleatorias independientes** de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 supuestas conocidas.

Un estimador puntual de la diferencia de medias $\mu_1 - \mu_2$ es la estadística $\overline{X}_1 - \overline{X}_2$ cuyo valor $\bar{x}_1 - \bar{x}_2$ es la estimación puntual.

Si las *dos poblaciones son normales,* entonces, $\overline{X}_1$ y $\overline{X}_2$ tienen distribuciones respectivas normal $N(\mu_1, \sigma_1^2/n_1)$ y $N(\mu_2, \sigma_2^2/n_2)$ (para $n_1 \geq 2$, y $n_2 \geq 2$). En consecuencia, la estadística $\overline{X}_1 - \overline{X}_2$ tiene distribución normal $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Si las dos *poblaciones no* son *normales* pero n_1 y n_2 son suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$), entonces, la estadística $\overline{X}_1 - \overline{X}_2$ es aproximadamente normal $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Por tanto, según sea el caso, la variable aleatoria:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

tiene distribución exactamente o aproximadamente normal $N(0,1)$.

Dado el grado de confianza $1-\alpha$, en la distribución de Z se puede encontrar el valor $z_0 = z_{1-\alpha/2}$ tal que $P[-z_0 \leq Z \leq z_0] = 1-\alpha$. (fig. 9.5)

Sustituyendo $Z = (\overline{X}_1 - \overline{X}_2)/\sigma_{\overline{X}_1 - \overline{X}_2}$ en la probabilidad, donde el error estándar de la diferencia es $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ y operando, resulta,

$$P\left[(\overline{X}_1 - \overline{X}_2) - z_0\sigma_{\overline{X}_1-\overline{X}_2} \le \mu_1 - \mu_2 \le (\overline{X}_1 - \overline{X}_2) + z_0\sigma_{\overline{X}_1-\overline{X}_2}\right] = 1-\alpha.$$

Luego,

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias que resultan de dos muestras aleatorias independientes de tamaños n_1 y n_2 escogidas respectivamente de dos poblaciones con varianzas σ_1^2 y σ_2^2 supuestas conocidas, entonces, el intervalo de confianza del $(1-\alpha)100\%$ de $\mu_1 - \mu_2$ es:

$$(\bar{x}_1 - \bar{x}_2) - z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

El valor $z_0 = z_{1-\alpha/2}$ se obtiene de la tabla normal $N(0,1)$ de manera que:

$$P[Z \le z_{1-\alpha/2}] = 1 - \alpha/2\ .$$

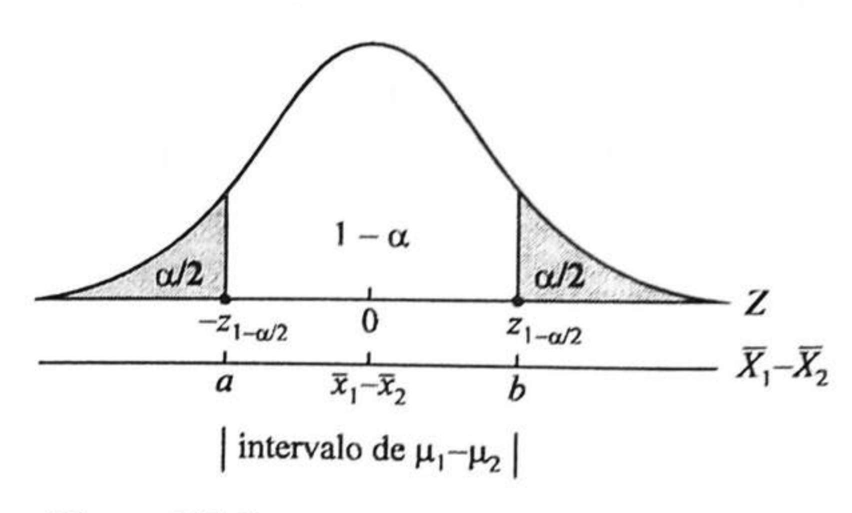


Figura 9.5: Intervalo de estimación de $\mu_1 - \mu_2$

La ilustración del intervalo de confianza de $\mu_1 - \mu_2$ es la figura 9.5, donde,

$$a = (\bar{x}_1 - \bar{x}_2) - z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \quad \text{y}$$

$$b = (\bar{x}_1 - \bar{x}_2) + z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

son los límites de confianza inferior y superior respectivamente de la diferencia de las medias poblacionales $\mu_1 - \mu_2$.

EJEMPLO 9.12.

Un ingeniero industrial a cargo de una planta de producción quiere determinar si hay diferencia en el número de unidades producidas en los dos turnos: matutino y vespertino. El quiere usar la técnica de estimar la diferencia entre las medias de los dos turnos sabiendo que las desviaciones estándares de las dos poblaciones son 44 y 65 respectivamente. Para esto escogió dos muestras aleatorias independientes de 55 obreros del turno matutino y 65 de obreros del turno vespertino de un día cualquiera, y halló que el número medio de unidades producidas respectivas fue de 435 y 400.

Determine un intervalo de confianza del 98% para la diferencia $\mu_1 - \mu_2$, donde μ_1 es la media de toda la producción matutina y μ_2 es la media de toda la producción vespertina.

¿Es acertada la decisión del ingeniero si infiere que no hay diferencia significativa en la producción de los dos turnos?

SOLUCION.

La estimación puntual de la diferencia de las dos medias poblacionales $\mu_1 - \mu_2$ es la diferencia de las medias muestrales:

$$\bar{x}_1 - \bar{x}_2 = 435 - 400 = 35$$

El error estándar de la diferencia de las dos medias es:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(44)^2}{55} + \frac{(65)^2}{65}} = 10.01$$

Para el nivel de confianza del 98% se encuentra:

$$z_0 = z_{1-\alpha/2} = z_{0.99} = 2.33.$$

Los límites de confianza inferior y superior respectivamente de $\mu_1 - \mu_2$ son

$$(\bar{x}_1 - \bar{x}_2) \mp z_0 \sigma_{\bar{x}_1 - \bar{x}_2} = 35 \mp 2.33 \times 10.01 = 35 \mp 23.3233$$

Luego, el intervalo de confianza aproximado del 98% para $\mu_1 - \mu_2$ es:

$$11.6767 < \mu_1 - \mu_2 < 58.3233$$

Dado que $\mu_1 - \mu_2 = 0 \notin [11.6767,\ 58.3233]$, se concluye que $\mu_1 \neq \mu_2$. Además, la diferencia $\mu_1 - \mu_2$ es positiva. Por lo tanto, se infiere que $\mu_1 > \mu_2$, es decir el turno matutino produce más.

9.7.2 Intervalo de confianza para la diferencia entre dos medias: Varianzas σ_1^2 y σ_2^2 supuestas desconocidas

A) Poblaciones no normales

Si $\bar{x}_1$ y $\bar{x}_2$ son los valores de las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones cuyas *distribuciones son no normales* con varianzas σ_1^2 y σ_2^2 supuestas desconocidas, entonces, siempre que los tamaños de las muestras sean grandes ($n_1 \geq 30$ y $n_2 \geq 30$), los parámetros σ_1 y σ_2 se estiman puntualmente por $\hat{s}_1$ y $\hat{s}_2$.

El intervalo de confianza del $(1-\alpha)\times 100\%$ para $\mu_1 - \mu_2$ es entonces:

$$(\bar{x}_1 - \bar{x}_2) - z_0\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_0\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}.$$

B) Poblaciones normales

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias y $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 respectivamente seleccionadas de *dos poblaciones normales* con varianzas σ_1^2 y σ_2^2 supuestas desconocidas.

B1) Varianzas supuestas iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

En este caso, las variables aleatorias,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}, \quad \text{y}, \quad V = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{\sigma^2},$$

tienen respectivamente distribución normal $N(0,1)$, y chi-cuadrado con grados de libertad: $n_1 + n_2 - 2$. Además, se comprueba que ambas U y V son independientes. Entonces, la variable aleatoria:

$$T = \frac{U}{\sqrt{\dfrac{V}{n_1 + n_2 - 2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\hat{S}_c^2}{n_1} + \dfrac{\hat{S}_c^2}{n_2}}}$$

tiene distribución *t*-student con $n_1 + n_2 - 2$ grados de libertad.

La varianza común muestral $\hat{S}_c^2$ definida por:

$$\hat{S}_c^2 = \frac{(n_1-1)\hat{S}_1^2 + (n_2-1)\hat{S}_2^2}{n_1+n_2-2}$$

es un *estimador insesgado de la varianza común poblacional* σ^2.

Dado el nivel de confianza $1-\alpha$, en la distribución $t(n_1+n_2-2)$ se halla el valor $t_0 = t_{1-\alpha/2, n_1+n_2-2}$ (figura 9.6), tal que: $P[-t_0 \le T \le t_0] = 1-\alpha$.

Sustituyendo la expresión de T, y manipulando algebraicamente, resulta,

$$P\left[(\bar{X}_1-\bar{X}_2) - t_0\sqrt{\hat{S}_c^2/n_1 + \hat{S}_c^2/n_2} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2) + t_0\sqrt{\hat{S}_c^2/n_1 + \hat{S}_c^2/n_2}\right] = 1-\alpha.$$

Luego,

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias que resultan de dos muestras independientes de tamaños n_1 y n_2 escogidas respectivamente de dos poblaciones normales con varianzas σ_1^2 y σ_2^2 supuestas desconocidas e iguales, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ de $\mu_1-\mu_2$ es: $$(\bar{x}_1-\bar{x}_2) - t_0 \times ES \le \mu_1-\mu_2 \le (\bar{x}_1-\bar{x}_2) + t_0 \times ES$$

donde $ES = \sqrt{\hat{s}_c^2/n_1 + \hat{s}_c^2/n_2}$ es el valor del error estándar de la diferencia de medias

El valor $t_0 = t_{1-\alpha/2, n_1+n_2-2}$ se busca en la tabla t-student con n_1+n_2-2 grados de libertad tal que: $P[T \le t_{1-\alpha/2, n_1+n_2-2}] = 1-\alpha/2$.

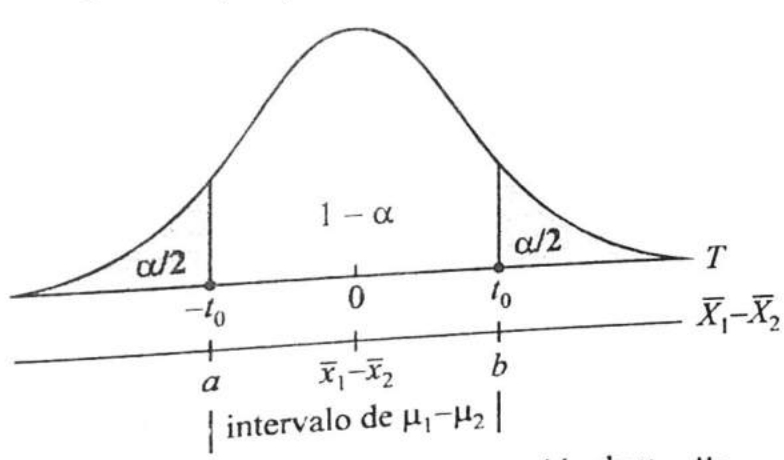


Figura 9.6: Intervalo de estimación de $\mu_1-\mu_2$

La ilustración es la figura 9.6, donde:

$$a = (\bar{x}_1-\bar{x}_2) - t_0 \times ES \quad \text{y} \quad b = (\bar{x}_1-\bar{x}_2) + t_0 \times ES$$

son los límites de confianza de $\mu_1-\mu_2$, inferior y superior respectivamente

B2) Varianzas supuestas distintas: $\sigma_1^2 \neq \sigma_2^2$

En este caso, la estadística,

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}}$$

tiene distribución *t*-student con *r* grados de libertad, siendo,

$$r = \frac{\left[\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right]^2}{\frac{\left[\frac{\hat{s}_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{\hat{s}_2^2}{n_2}\right]^2}{n_2 - 1}}.$$

Dado que r rara vez es un entero, se redondea al entero más cercano.

El intervalo de confianza del $(1-\alpha)\times 100\%$ para $\mu_1 - \mu_2$ se obtiene de los limites de confianza inferior y superior respectivamente:

$$\left(\bar{x}_1 - \bar{x}_2\right) - t_{1-\alpha/2,\, r} \times ES \quad \text{y} \quad \left(\bar{x}_1 - \bar{x}_2\right) + t_{1-\alpha/2,\, r} \times ES$$

donde: $ES = \sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}$ es el error estándar de la estadística diferencia de medias y el valor $t_{1-\alpha/2,\, r}$ se encuentra en la tabla t-student con r grados de libertad tal que: $P[T \leq t_{1-\alpha/2}] = 1 - \alpha/2$.

EJEMPLO 9.13.

Una compañía exportadora de café quiere escoger la mejor calidad de café de exportación entre dos variedades de café en grano: A (Chanchamayo) y B (Quillabamba). Elegirá la variedad de café que contenga el menor porcentaje de impurezas por saco de un quintal. Se sabe que los porcentajes de impurezas por saco de cada variedad de café tienen distribución normal y con la misma varianza.

Dos muestras aleatorias independientes una de 10 sacos de A y la otra de 12 sacos de B, revelaron los siguientes porcentajes de impurezas por saco de café.

A: 4, 3, 6, 6, 5, 6, 7, 4, 7, 6.
B: 7, 6, 10, 8, 9, 8, 7, 6, 7, 9, 5, 8.

Estime mediante un intervalo de confianza del 95% la diferencia entre los dos promedios de porcentajes de impurezas por saco de toda la producción de las dos variedades de café, ¿qué variedad de café debería elegir para la exportación?.

SOLUCION.

Sean X_1 y X_2 las poblaciones de porcentajes de impurezas por saco de café de A y B respectivamente. Se supone que **las poblaciones son normales con varianzas desconocidas supuestas iguales.**

De las muestras se obtiene:

$$n_1 = 10, \quad \bar{x}_1 = 5.4, \quad \hat{s}_1 = 1.3499,$$

$$n_2 = 12, \quad \bar{x}_2 = 7.5, \quad \hat{s}_2 = 1.446,$$

$$\hat{s}_c^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2} = \frac{9(1.3499)^2 + 11(1.446)^2}{10+12-2} = 1.97$$

El **error estándar** estimado de la diferencia de medias (*ES*) es el número:

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\hat{s}_c^2}{n_1} + \frac{\hat{s}_c^2}{n_2}} = \sqrt{\frac{1.97}{10} + \frac{1.97}{12}} = 0.601.$$

Para $1-\alpha = 0.95$ y 20 grados de libertad se halla: $t_{0.975,\,20} = 2.086$

Los límites de confianza inferior y superior del 95% para $\mu_1 - \mu_2$ son:

$$(\bar{x}_1 - \bar{x}_2) \mp t_{0.975,\,20} \times ES = (5.4 - 7.5) \mp 2.086 \times 0.601 = -2.1 \mp 1.254.$$

Luego, el intervalo de confianza del 95% para $\mu_1 - \mu_2$ es

$$-3.354 < \mu_1 - \mu_2 < -0.846$$

Como el valor: $\mu_1 - \mu_2 = 0$ no pertenece al intervalo de confianza, no se debería aceptar que: $\mu_1 = \mu_2$. Además, como $\mu_1 - \mu_2 < 0$, debería exportar la variedad A.

EJEMPLO 9.14.

Se lleva a cabo un estudio para comparar el sueldo en Lima de los ingenieros (B1) y de los administradores (B2) egresados de la PUCP. Las experiencias anteriores indican que la distribución de los sueldos tanto en B1 como en B2 es normal y con varianzas diferentes. Dos muestras aleatorias, una de 9 sueldos de B1 y otra de 8 sueldos de B2 dieron los siguientes ingresos en miles de dólares:

B1: 1.2, 2.8, 1.0, 2.5, 2.4, 1.9, 2.2, 3.3, 1.7
B2: 1.6, 2.0, 1.6, 2.0, 1.6, 1.7, 1.5, 2.1.

Utilizando un intervalo de confianza del 95% para la verdadera diferencia de las dos medias de los sueldos, ¿se puede concluir que los ingenieros ganan menos que los administradores, egresados de la PUCP?.

SOLUCION.

Sean X_1 y X_2 las variables aleatorias que representan los sueldos de B1 y B2 respectivamente. Se supone varianzas poblacionales diferentes.

De los datos de las muestras se obtiene:

$$n_1 = 9, \quad \bar{x}_1 = 2, \quad \hat{s}_1 = 0.604$$
$$n_2 = 8, \quad \bar{x}_2 = 1.763, \quad \hat{s}_2 = 0.233$$

La diferencia de las medias muestrales es: $\bar{x}_1 - \bar{x}_2 = 0.237$

El error estándar de la diferencia de medias *ES* es el número:

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = \sqrt{\frac{(0.604)^2}{9} + \frac{(0.233)^2}{8}} = 0.218$$

El número de grados de libertad es:

$$g = \frac{\left[\hat{s}_1^2/n_2 + \hat{s}_2^2/n_2\right]^2}{\frac{\left[\hat{s}_1^2/n_1\right]^2}{n_1 - 1} + \frac{\left[\hat{s}_2^2/n_2\right]^2}{n_2 - 1}} = \frac{\left[(0.604)^2/9 + (0.233)^2/8\right]^2}{\frac{\left[(0.604)^2/9\right]^2}{9-1} + \frac{\left[(0.233)^2/8\right]^2}{8-1}} = 10.56 \cong 11$$

Para $1 - \alpha = 0.95$ y $r = 11$ grados de libertad se obtiene $t_{0.975, 11} = 2.201$.

Los límites de confianza inferior y superior aproximados de $\mu_1 - \mu_2$ son

$$(\bar{x}_1 - \bar{x}_2) \mp t_{0.975, 11} \times ES = 0.237 \mp 2.201 \times 0.218 = 0.237 \mp 0.4798$$

Luego, el intervalo de confianza del 95 por 100 para $\mu_1 - \mu_2$ es

$$-0.2428 < \mu_1 - \mu_2 < 0.7168$$

Dado que $\mu_1 - \mu_2 = 0 \in [-0.2428, \ 0.7168]$, podemos concluir que $\mu_1 = \mu_2$. Por lo tanto, no hay diferencias significativas entre las medias de los sueldos de los egresados de Ingeniería y Administración de la PUCP.

NOTA El lector debería resolver los ejemplos 9.13 y 9.14 utilizando un paquete de computo. Por ejemplo, el *MCEST*.

9.8 Intervalo de confianza para la diferencia entre dos medias con observaciones pareadas

Sea $(X_1,Y_1), (X_2,Y_2),\ldots, (X_n,Y_n)$ una muestra aleatoria de n datos aparejados, donde las muestras $X_1, X_2,\ldots, X_n$ e $Y_1, Y_2,\ldots, Y_n$ **correlacionadas**, son seleccionadas respectivamente de dos poblaciones normales $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$.

Podemos concebir estas n diferencias: $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2,\ldots,$ $D_n = X_n - Y_n$ como una muestra aleatoria seleccionada de una población de diferencias $D = X - Y$ cuya distribución es normal $N(\mu_D, \sigma_D^2)$, con media $\mu_D = \mu_1 - \mu_2$ y varianza $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\operatorname{cov}(X,Y)$.

El estimador insesgado de $\mu_D = \mu_1 - \mu_2$ es la estadística:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{\sum_{i=1}^{n}(X_i - Y_i)}{n} = \frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{X} - \overline{Y}$$

cuyo valor $\overline{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$ es la estimación insesgada de μ_D.

Si σ_D^2 es conocida, la estadística $\overline{D}$ tiene distribución normal $N(\mu_D, \sigma_D^2/n)$. Consecuentemente la estadística:

$$Z = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sigma_D/\sqrt{n}}$$

tiene distribución normal $N(0,1)$.
En esta distribución se determina el intervalo e confianza de: $\mu_D = \mu_1 - \mu_2$.

Luego, si $\overline{d}$ es la media de las diferencias de n datos pareados escogidos de una distribución normal con desviación estándar σ_D supuesta conocida, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ para la diferencia de medias $\mu_D = \mu_1 - \mu_2$ es la expresión:

$$\bar{d} - z_{1-\alpha/2}\,\sigma_D/\sqrt{n} \le \mu_1 - \mu_2 \le \bar{d} + z_{1-\alpha/2}\,\sigma_D/\sqrt{n}$$

El valor $z_{1-\alpha/2}$ se encuentra en la tabla normal $N((0,1)$, tal que

$$P[Z \le z_{1-\alpha/2}] = 1 - \alpha/2$$

Por otra parte, si la varianza σ_D^2 es *desconocida*, el estimador puntual de esta varianza, es la estadística:

$$\hat{S}_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$$

cuyo valor, $$\hat{s}_d^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = \frac{\sum_{i=1}^{n}d_i^2 - n(\bar{d})^2}{n-1}$$

es la estimación de σ_D^2. Además se verifica que la estadística

$$T = \frac{\bar{D} - (\mu_1 - \mu_2)}{\hat{S}_D/\sqrt{n}}$$

tiene distribución t-student con $n-1$ grados de libertad. Esta estadística se usa para determinar el intervalo de confianza de $\mu_D = \mu_1 - \mu_2$

Luego,

Si $\bar{d}$ y $\hat{s}_d$ son la media y la desviación estándar de una muestra aleatoria de n diferencias de pares de datos de una población normal con varianza σ_D^2 *supuesta desconocida*, entonces el intervalo de confianza del $(1-\alpha)\times 100\%$ para $\mu_D = \mu_1 - \mu_2$ es

$$\bar{d} - t_{1-\alpha/2,\,n-1} \times ES \le \mu_1 - \mu_2 \le \bar{d} + t_{1-\alpha/2,\,n-1} \times ES$$

donde: $ES = \hat{s}_d/\sqrt{n}$ es el error estándar y el valor $t_{1-\alpha/2,\,n-1}$ se encuentra en la tabla t-student con $n-1$ grados de libertad tal que: $P[T \le t_{1-\alpha/2,\,n-1}] = 1 - \alpha/2$.

EJEMPLO 9.15.

El gerente de operaciones de una empresa de confecciones debe tomar la decisión entre dos procesos de manufactura A y B para la fabricación de una prenda. Eligió 10 operadores eficientes y cada uno de ellos utilizó los dos procesos de manufactura para fabricar la prenda, resultando los siguientes tiempos en minutos para los procesos de manufactura A y B.

Operador	1	2	3	4	5	6	7	8	9	10
Proceso A	10	10	11	12	12	13	13	15	15	16
Proceso B	8	11	11	10	9	11	10	12	14	15
Diferencia d_i	2	−1	0	2	3	2	3	3	1	1
d_i^2	4	1	0	4	9	4	9	9	1	1

Utilizando un intervalo de confianza del 98 %, ¿se puede afirmar que el proceso de manufactura B reduce el tiempo medio de fabricación de la prenda?. Suponga distribución normal de la diferencia de los tiempos.

SOLUCION.

Sean X_i, Y_i los tiempos de los procesos A y B respectivamente

Haciendo: $d_i = X_i - Y_i$, $d_i = 1, 2, ..., 10$ se obtiene,

$$\bar{d} = \frac{\sum_{i=1}^{10} d_i}{10} = \frac{16}{10} = 1.6.$$

$$\hat{s}_d = \sqrt{\frac{\sum_{i=1}^{10}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{42 - 10(1.6)^2}{10-1}} = 1.3499.$$

Error estándar es: $ES = \hat{s}_d / \sqrt{n} = 1.3499/\sqrt{10} = 0.42688.$

Para $1-\alpha = 0.98$ y para los grados de libertad: $n-1=9$ en la tabla t-Student se halla: $t_{1-\alpha/2, n-1} = t_{0.99, 9} = 2.821$.

Los límites de confianza para $\mu_D = \mu_1 - \mu_2$ son,

$$\bar{d} \mp t_{1-\alpha/2, n-1} \times ES = 1.6 \mp 2.821 \times 0.42688 = 1.6 \mp 1.2042.$$

Por tanto: $\mu_D = \mu_1 - \mu_2 \in [0.3958,\ 2.8042]$ con confianza del 95%

Este intervalo no contiene el valor $\mu_D = \mu_1 - \mu_2 = 0$, entonces, debemos afirmar que las medias de ambos tiempos son diferentes. Además como $\mu_1 - \mu_2 > 0$, el proceso de manufactura B reduce el tiempo de fabricación de la prenda.

NOTA. Utilizando el paquete de computo ***MCEST*** con confianza al 98 % se tiene también $\mu_D \in [0.3958,\ 2.8042]$.

9.9 Intervalo de confianza para una proporción

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n escogida de una población de Bernoulli $B(1, p)$, cuyo parámetro p es la ***proporción de éxitos en la población***.

En la muestra cada $X_i = 1$ si hay éxito con probabilidad: p, y cada $X_i = 0$ si no hay éxito con probabilidad $1 - p$.

El estimador puntual de p es la estadística ***proporción de éxitos en la muestra*** definida por:

$$\overline{P} = \frac{X}{n}$$

en donde X es el número de éxitos en la muestra: $X = \sum_{i=1}^{n} X_i$ cuya distribución de probabilidades es Binomial $B(n, p)$.

El valor $\overline{p} = x/n$ que se obtiene de una muestra, es la estimación puntual del parámetro p.

Si n es suficientemente grande ($n \geq 30$), la distribución de probabilidad de la proporción estadística $\overline{P}$ es aproximadamente normal con media: $\mu_{\overline{P}} = E(\overline{P}) = p$ y con varianza: $\sigma^2_{\overline{P}} = Var(\overline{P}) = p(1-p)/n$.

En consecuencia, la variable aleatoria

$$Z = \frac{\overline{P} - p}{\sqrt{p(1-p)/n}}$$

tiene distribución aproximadamente $N(0,1)$.

Esta distribución se utiliza para determinar un intervalo de confianza para el parámetro p.

Dado el nivel de confianza $1 - \alpha$, en la distribución de Z, se halla el valor $z_{1-\alpha/2}$ tal que $P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1 - \alpha$. (ver figura 9.7)

Sustituyendo la expresión de Z resulta

$$P[-z_{1-\alpha/2} \leq \frac{\overline{P} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}] = 1 - \alpha.$$

$$P[\overline{P} - z_{1-\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \overline{P} + z_{1-\alpha/2}\sqrt{p(1-p)/n}] = 1 - \alpha.$$

El error estándar de la estadística $\overline{P}$ es: $\sigma_{\overline{P}} = \sqrt{p(1-p)/n}$ cuya estimación es el valor: $\hat{\sigma}_{\overline{P}} = \sqrt{\overline{p}(1-\overline{p})/n}$.

Luego,

Si $\overline{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ para p es:

$$\overline{p} - z_{1-\alpha/2} \times \sqrt{\overline{p}(1-\overline{p})/n} \leq p \leq \overline{p} + z_{1-\alpha/2} \times \sqrt{\overline{p}(1-\overline{p})/n}$$

La ilustración es la figura 9.7, donde

$$a = \overline{p} - z_{1-\alpha/2} \times ES \quad \text{y} \quad b = \overline{p} + z_{1-\alpha/2} \times ES$$

son los límites de confianza de p, inferior y superior respectivamente, y donde además *ES* es el valor del error estándar

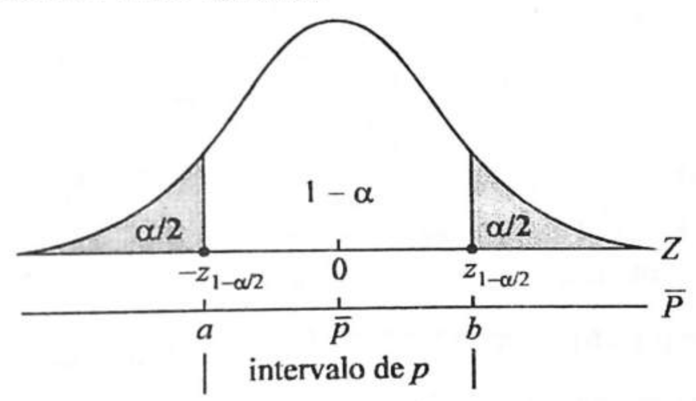


Figura 9.7: Intervalo de estimación del $(1-\alpha)\times 100\%$ para p

EJEMPLO 9.16.

Se van a realizar las elecciones para elegir el alcalde de Lima. La campaña del candidato A y su oponente es muy candente. En una reciente encuesta el candidato A se ubicaba en segundo lugar. Para comprobar este resultado, el personal de campaña de A va a realizar una encuesta propia la última semana permitida por la ley electoral.

a) ¿Qué tamaño de muestra deberían seleccionar si desean tener una confianza del 95% de que el error de estimación de la proporción a favor de A en toda la población de electores no sea superior a 5% y si se dispone de una estimación para la proporción de la población a favor de A en 40%?.

b) ¿Cuál es el tamaño de la muestra que deberían escoger si desean tener una confianza del 95% de que el error de estimación de la proporción a favor de A

en toda la población de electores no sea superior a 5% y si no se dispone de una estimación para la proporción de la población?.

c) El personal de campaña de A realizó una encuesta con el tamaño de muestra calculada en b). Si la muestra reveló que 154 votarían a favor de A, estime el porcentaje de electores a favor de A en toda la población, utilizando un intervalo de confianza del 95%. Interprete el resultado

d) Si con la misma muestra que escogió el personal de campaña de A, ellos estiman que la proporción a favor de su oponente es 42%, con una confianza del 95% y un error de estimación igual a 5%, ¿pueden ellos concluir que el candidato A perdería definitivamente las elecciones?.

SOLUCION:

a) Si se utiliza el valor previo $\overline{p}$ de una *muestra preliminar o piloto*, o una estimación anterior del parámetro, el error máximo de estimación de p es:

$$e = z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n}$$

de donde resulta; $$n = \frac{(z_{1-\alpha/2})^2\,\overline{p}(1-\overline{p})}{e^2}.$$

Para $1-\alpha = 0.95$, se obtiene: $z_{1-\alpha/2} = z_{0.975} = 1.96$. Luego, se tiene una confianza del 95% que el error al estimar p no será mayor que 0.05 si el tamaño de la muestra es al menos,

$$n = [(1.96)^2 \times 0.4 \times 0.6]/(0.05)^2 = 368.7936 \cong 369$$

b) Si no se tiene una previa estimación puntual del parámetro p, entonces, se hace $\overline{p}(1-\overline{p}) = 1/4$ (el valor máximo que pueden asumir ambos $\overline{p}$ y $1-\overline{p}$ es 1/4 o también se verifica que $\overline{p}(1-\overline{p}) = -(\overline{p}-1/2)^2 + 1/4 \leq 1/4$). Entonces,

$$n \cong \frac{(z_{1-\alpha/2})^2}{4e^2}.$$

Para $1-\alpha = 0.95$ resulta: $z_{1-\alpha/2} = z_{0.975} = 1.96$. Luego, se tiene una confianza del 95% que el error al estimar p no será mayor que 0.05 si el tamaño de la muestra es al menos:

$$n = (1.96)^2/(4 \times (0.05)^2) = 384.16 \cong 385$$

c) La estimación puntual de la proporción p a favor de A en la población, es la proporción a su favor en la muestra de $n = 385$ electores; esto es,

$$\overline{p} = 154/385 = 0.40.$$

El error **estándar** estimado de la estadística $\overline{P}$ es el número:

$$\hat{\sigma}_{\overline{P}} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = \sqrt{\frac{(0.40)(0.60)}{385}} = 0.0249675.$$

Para $1-\alpha = 0.95$ se obtiene: $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Los límites de confianza de p, inferior y superior, son respectivamente

$$\overline{p} \mp z_{1-\alpha/2} \times \hat{\sigma}_{\overline{P}} = 0.40 \mp 1.96 \times 0.0249675 = 0.40 \mp 0.048936$$

Luego, el intervalo de confianza del 95% para p es de 0.3511 a 0.4489.

Es decir, $p \in [35.11\%, 44.89\%]$ con confianza del 95%. También, si la proporción favor de A en toda la población se estima en 40% se tiene una confianza del 95% que el error de la estimación no es superior a 0.048936.

d) El intervalo de confianza del 95% a favor de A es [35.11%, 44.89%].
El intervalo de confianza del 95% a favor de su oponente es [37%, 47%].
Dado que la intersección de los dos intervalos no es vacía, no se puede proclamar al oponente como ganador absoluto.
En este caso se dice que hay un **empate técnico**.

NOTA. Si el muestreo es sin reemplazo en una población finita de tamaño N, entonces el error estándar de $\overline{P}$ es: $\sigma_{\overline{P}} = \left(pq/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$. El error estándar estimado *ES* de la estadística es:

$$\hat{\sigma}_{\overline{P}} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\sqrt{\frac{N-n}{N-1}}.$$

y el valor de **n** se calcula por: $$n = \frac{z_{1-\alpha/2}^2\,\overline{pq}N}{z_{1-\alpha/2}^2\,\overline{pq} + e^2(N-1)}.$$

Si no se tiene el dato $\overline{p}$, se puede utilizar el valor $\overline{p}$=0.5.

NOTA. En estadística aplicada es frecuente utilizar: $z_{1-\alpha/2} = 2$. En este caso, el tamaño de la muestra requerida para estimar p está dada por la expresión:

$$n = \frac{N\overline{pq}}{\overline{pq} + \frac{e^2}{4}(N-1)}$$

Si no se tiene el dato $\overline{p}$, se puede utilizar el valor $\overline{p}$=0.5. (Ver problema 34)

Si el método de muestreo es por estratos, el tamaño de la muestra requerida para estimar p con un error máximo de estimación e está dado por:

$$n=\frac{\sum_{1}^{k} N_i^2 \overline{p}_i \overline{q}_i / w_i}{\frac{e^2}{4}N^2+\sum_{1}^{k} N_i \overline{p}_i \overline{q}_i}$$

donde: w_i es el % de observaciones asignados al estrato i, para $i=1,2,\ldots,k$.

Si se desconoce $\overline{p}_i$ se puede utilizar el valor $\overline{p}_i=0.5$.

EJEMPLO 9.17.

Se planea una investigación para determinar el porcentaje de turistas extranjeros que gastan mas de $100 diarios en el Cuzco. Se supone que diariamente el Cuzco atiende a 3000 turistas extranjeros.

a) ¿Qué tamaño de muestra debería escoger el investigador para realizar este estudio si quiere tener una confianza del 95% de que el error de la estimación de la proporción de turistas extranjeros que gastan mas de $100 diarios en el Cuzco no sea superior al 4%?.

b) Con una muestra del tamaño calculado en a) el investigador encontró que 150 de ellos gastaron más de $100 diarios. Además, decidió utilizar la proporción muestral $\overline{p}=0.30$ como estimación de la proporción de todos los turistas extranjeros que gastan mas de $100 diarios en el Cuzco. Si estimó de 780 a 1020 el total de turistas extranjeros que gastan mas de $100 diarios en el Cuzco, ¿qué grado de confianza utilizó?

SOLUCION.

a) Para $1-\alpha=0.95$, se obtiene: $z_{1-\alpha/2}=z_{0.975}=1.96$.

Utilizando el valor $\overline{p}(1-\overline{p})=1/4$ y $N=3000$ se tiene:

$$n=\frac{z_{1-\alpha/2}^2\overline{pq}N}{z_{1-\alpha/2}^2\overline{pq}+e^2(N-1)}=\frac{(1.96)^2(3000)}{(1.96)^2+4(0.04)^2(3000-1)}=500.3125\cong 501.$$

b) El intervalo: $780\le Np\le 1020$, resulta de: $N(\overline{p}\mp z_{1-\alpha/2}\sigma_{\overline{p}})$, donde::

$$\sigma_{\overline{p}}=\sqrt{(\overline{p}(1-\overline{p})/n)((N-n)/(N-1))}.$$

Para $n=589$, N=3000 y $\overline{p}=0.30$, se obtiene $\sigma_{\overline{p}}=$ **0.018688946.**

De $1020=N(\overline{p}+z_{1-\alpha/2}\sigma_{\overline{p}})$ resulta: $z_{1-\alpha/2}=2.17$, $1-\alpha/2=0.985$, $\alpha=0.03$, y $1-\alpha=0.97$

9.10 Intervalo de confianza para la diferencia entre dos proporciones

Sean $\overline{P}_1$ y $\overline{P}_2$ las proporciones de éxitos de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones de Bernoullí $B(1, p_1)$ y $B(1, p_2)$, donde p_1 y p_2 son los respectivos parámetros proporciones de éxitos.

La estimación puntual de $p_1 - p_2$ es la estadística $\overline{P}_1 - \overline{P}_2$ cuyo valor es $\overline{p}_1 - \overline{p}_2$ obtenida de las dos muestras.

Si n_1 y n_2 son suficientemente grandes, entonces, $\overline{P}_1$ y $\overline{P}_2$ tienen distribuciones aproximadamente normales respectivas $N(p_1, p_1(1-p_1)/n_1)$ y $N(p_2, p_2(1-p_2)/n_2)$.

Por lo tanto, por la propiedad reproductiva de la normal, la estadística $\overline{P}_1 - \overline{P}_2$ tendrá distribución aproximadamente normal
con media:

$$\mu_{\overline{P}_1 - \overline{P}_2} = E(\overline{P}_1 - \overline{P}_2) = p_1 - p_2,$$

y con varianza

$$\sigma^2_{\overline{P}_1 - \overline{P}_2} = Var(\overline{P}_1 - \overline{P}_2) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2.$$

Por consiguiente, la estadística:

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}},$$

tendrá distribución aproximadamente normal $N(0,1)$, siendo $q_1 = 1 - p_1$ y $q_2 = 1 - p_2$.

Esta distribución se utiliza para determinar el intervalo de confianza de $p_1 - p_2$.

Dado el nivel de confianza: $1 - \alpha$, en la distribución de Z se encuentra el valor $z_0 = z_{1-\alpha/2}$ tal que:

$$P[-z_0 \leq Z \leq z_0] = 1 - \alpha.$$

Sustituyendo el valor de Z y luego de hacer manipulaciones algebraicas, se obtiene:

$$P\left[(\overline{P}_1-\overline{P}_2)-z_0\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}\le p_1-p_2\le(\overline{P}_1-\overline{P}_2)+z_0\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}\right]=1-\alpha.$$

El error estándar de la estadística $\overline{P}_1-\overline{P}_2$ es :

$$\sigma_{\overline{P}_1-\overline{P}_2}=\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}},$$

Cuya estimación es el valor:

$$\hat{\sigma}_{\overline{P}_1-\overline{P}_2}=\sqrt{\frac{\overline{p}_1\overline{q}_1}{n_1}+\frac{\overline{p}_2\overline{q}_2}{n_2}},$$

que por comodidad será denotada por *ES*.

Luego,

Si $\overline{p}_1$ y $\overline{p}_2$ son las proporciones de éxitos en dos muestras aleatorias independientes de tamaños n_1 y n_2 respectivamente, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ del parámetro p_1-p_2 es

$$(\overline{p}_1-\overline{p}_2)-z_{1-\alpha/2}\times ES\le p_1-p_2\le(\overline{p}_1-\overline{p}_2)+z_{1-\alpha/2}\times ES$$

El valor $z_{1-\alpha/2}$ se halla en la tabla normal estándar, de manera que: $P[Z\le z_{1-\alpha/2}]=1-\alpha/2$.

EJEMPLO 9.18.

El fabricante de la cerveza "Dorada" afirma que su marca de cerveza es más preferida en Iquitos que en Cuzco. Para comprobar esta afirmación un investigador de mercado escogió dos muestras aleatorias, una de 500 consumidores de cerveza en Iquitos y otra de 400 consumidores de cerveza en el Cuzco. Si las muestras revelaron que 350 consumidores en Cuzco y 240 consumidores en Iquitos prefieren la cerveza "DORADA", utilizando un intervalo de confianza del 95% para la diferencia de dos proporciones de todos los consumidores de esta cerveza en las dos ciudades, ¿se puede inferir que el fabricante tiene la razón?

SOLUCION.

Sean $\overline{p}_1$ y $\overline{p}_2$ las proporciones muestrales que consumen la cerveza "Dorada" en Iquitos y Cuzco respectivamente.

De los datos del problema se obtiene:

$$\bar{p}_1 = 350/500 = 0.7,$$

$$\bar{p}_2 = 240/400 = 0.6,$$

$$\bar{p}_1 - \bar{p}_2 = 0.7 - 0.6 = 0.1$$

La estimación puntual del parámetro diferencia de proporciones poblacionales: $p_1 - p_2$, es la diferencia de proporciones muestrales: $\bar{p}_1 - \bar{p}_2 = 0.1$

El error estándar *ES* de la estadística diferencia de proporciones muestrales es:

$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} = \sqrt{\frac{0.7 \times 0.3}{500} + \frac{0.6 \times 0.4}{400}} = 0.03194$$

Para el nivel de confianza: $1 - \alpha = 0.95$, se obtiene: $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Los límites de confianza del intervalo de $p_1 - p_2$ son:

$$(\bar{p}_1 - \bar{p}_2) \mp z_{1-\alpha/2} \times ES = 0.1 \mp 1.96 \times 0.03194 = 0.1 \mp 0.0626.$$

En consecuencia, el intervalo de confianza del 95% para $p_1 - p_2$ es

$$0.0374 \leq p_1 - p_2 \leq 0.1626.$$

Dado que el intervalo resultante no contiene el valor cero, debemos concluir que las proporciones de preferencias en la población son diferentes, esto es, $p_1 \neq p_2$. Además, dado que el intervalo contiene valores positivos, hay razones suficientes para concluir que $p_1 > p_2$.

EJERCICIOS

Adicionales de estimación puntual

1. La duración, en horas, de cierta clase de foco sigue una distribución exponencial con media desconocida θ horas. Se toma una muestra de un sólo foco al azar y se mide su duración X en horas.
   a) Si con X se estima θ, ¿se podría decir que X es un estimador insesgado del parámetro θ?.
   b) Si a θ se le estima mediante el intervalo de confianza, $\left[\frac{X}{20}, 20X\right]$, ¿cuál es el nivel de confianza de este intervalo?.

   Rp. a) $X \sim \text{Exp}(1/\theta)$, $E(X) = (1/\theta)^{-1} = \theta$. Si, X es un estimador insesgado de θ.,
   b) $1-\alpha=P[X/20<\theta<20X]=P[X<20\theta \wedge X>\theta/20]=e^{-1/20}-e^{-20}=0.951$.

2. El error de medición de un instrumento se asume que es una variable aleatoria con distribución normal $N(\mu,\sigma^2)$. Se dispone de una muestra aleatoria $X_1, X_2, ..., X_{10}$ de 10 mediciones con este instrumento.

   Para estimar μ el ingeniero A usa el estimador: $\hat{\theta}_1 = \frac{\sum_{i=1}^{10} X_i}{10}$

   Mientras que el ingeniero B usa el estimador: $\hat{\theta}_2 = \frac{\left(\sum_{i=1}^{10} X_i\right) - X_6}{9}$

   En términos de las propiedades de insegabilidad y varianza de los estimadores, ¿a quién le daría usted la razón?.

   Rp. a) $E(\hat{\theta}_1) = \mu$, $E(\hat{\theta}_2) = \mu$, b) $V(\hat{\theta}_1) = \sigma^2/10$, $V(\hat{\theta}_2) = 11\sigma^2/81$. la razón para A.

3. Sea X_1, X_2, X_3, X_4, X_5 una muestra aleatoria de una población $X \sim N(\mu, 1)$. Para estimar μ se proponen los siguientes estimadores:

   $$\hat{\theta}_1 = \frac{X_1 + X_5}{2} \quad \text{y} \quad \hat{\theta}_2 = \frac{X_1 + X_2 + X_3 + X_4 + 1}{4}$$

   a) ¿Son estos estimadores insesgados?
   b) Si se desea tener una alta probabilidad de que el estimador difiera de μ en no más de una unidad, ¿cuál de los dos estimadores anteriores escogería?

   Rp. a) $E(\hat{\theta}_1)=\mu$, $E(\hat{\theta}_2)=\mu+1/4$. b) $P[|\hat{\theta}_1-\mu|\leq 1]=P[|Z|\leq 2^{1/2}]=0.8429$.
   $P[|\hat{\theta}_2-\mu|\leq 1]=P[-5/2\leq Z\leq 3/2]=0.927$.

4. El tiempo (en horas) que esperan los pasajeros para abordar un avión se distribuye uniformemente en el intervalo $[0, \theta]$. Para estimar θ se seleccionan al azar a dos pasajeros que esperan subir a un avión y se observan los tiempos X_1 y X_2 hasta abordar al avión. Si se utilizan los siguientes estimadores de θ:

$$\hat{\theta}_1 = 2X_1 \quad \text{y} \quad \hat{\theta}_2 = X_1 + X_2$$

a) ¿Son tales estimadores insesgados?
b) ¿Cuál de los estimadores escogería como el mejor y por qué?
c) ¿Con qué probabilidad $\hat{\theta}_1$ difiere de θ en más de 0.25 horas si es que θ es igual a 2 horas?

Rp. a) $E(\hat{\theta}_1)=\theta$, $E(\hat{\theta}_2)=\theta$ son insesgados, b) $V(\hat{\theta}_1)=\theta^2/3$, $V(\hat{\theta}_2)=\theta^2/6$, la segunda es menor, c) $P[|\hat{\theta}_1-\theta|>0.25]=1-P[1.75\leq X_1\leq 2.25]=0.75$.

Una media

1. Un grupo de inversionistas quiere estimar la media del rendimiento anual de ciertos valores. Seleccionó una muestra aleatoria de 49 de los valores observando una media 8.71% y una desviación estándar 2.1%.
   a) Determine el error estándar de la media
   b) Estime la media del rendimiento anual de tales valores mediante un intervalo de confianza del 96%. ¿Es el error de estimación inferior al 5%?
   c) Determine el nivel de confianza si el rendimiento anual medio de la población de estos valores se estima entre 7.96% y 9.46%.

Rp. a) 0.3, b) 8.71 ∓ 0.615, NO, c) $\alpha=0.0124$, $1-\alpha=0.9876$

2. Se tomó una muestra aleatoria de 9 clientes de un banco local para estimar la media del tiempo que utilizan en sus distintas operaciones. La media calculada en la muestra fue 9 minutos. Se sabe que la población de los tiempo es normal con una desviación estándar de 3 minutos.
   a) Halle los limites de confianza inferior y superior para la media de todos los tiempos, al nivel de confianza 0.97.
   b) Si la media de la población de todos los tiempos de las operaciones se estima en el intervalo de 7 a 13 minutos, ¿es el nivel de confianza mayor que 0.97?.

Rp. a) 6.83 y 11,17 b) 0.9544, No.

3. Se quiere estimar la media del nivel de ansiedad de todos los estudiantes preuniversitarios. Para esto se seleccionó una muestra aleatoria de tamaño 100

estudiantes preuniversitarios y se les planteó la prueba para medir la ansiedad, resultando una media de 70 puntos y una desviación estándar de 10 puntos.

a) ¿Cuánto es la estimación puntual para la media del nivel de ansiedad de la población?.
b) ¿Es el error de esta estimación puntual superior a 5 puntos, con nivel de confianza de 0.98?
c) Determine el intervalo de confianza del 95% para la media del nivel de ansiedad de todos los estudiantes preuniversitarios.
d) Si usted considera que el intervalo encontrado en c) no es muy preciso, ¿qué acción debería tomar para que el intervalo de estimación al 95% sea más preciso?.

Rp. a) 70, b) No, es 2.33, c)70 ∓ 1.96, No, es =2.33, c) aumentar el tamaño de la muestra.

4. Un estudiante de estadística aplicada quiere confirmar el peso neto medio de las latas de néctar de fruta con la etiqueta "19 onzas". El sabe que la población de los pesos netos es normal con una desviación estándar de 2 onzas.
   a) ¿Qué tamaño de muestra debería escoger si quiere estimar la media de la población de los pesos con error de 0.98, y nivel confianza del 0.95?.
   b) El seleccionó muestra aleatoria de 20 latas y obtuvo una media de 18.5 onzas. Desarrolle un intervalo de confianza del 95% para la media de la población de los pesos. Con este resultado, ¿Se aclaró la duda del estudiante?
   c) ¿Qué porcentaje de tales intervalos no contendrían la media de la población?

   Rp. a) $n \geq 16$, b) 18.5 ∓ 0.876, Si. c) 5% Por definición

5. El trabajo encargado a un grupo de estudiantes de estadística consiste en realizar una encuesta para estimar el tiempo promedio por semana que los niños de entre 2 y 10 años, ven televisión. La estimación debe quedar dentro de un rango de una hora, con confianza de 95%.
   a) ¿Qué tamaño de muestra deberían escoger si se sabe que la desviación estándar de la población es 3 horas?.
   b) ¿Qué tamaño de muestra deberían escoger si se sabe que el tiempo mínimo es 1hora y el tiempo máximo es 10 horas?.

   Rp. a) $n \cong 138.3 \cong 139$, b) $n \cong 34.57 \cong 35$

6. Un inversionista analiza un informe sobre ingresos familiares en la región de San martín. El informe dice entre otras cosas que la población de los ingresos es normal con media de \$400 y que el 95% de todos los ingresos familiares mensuales varían de \$302 a \$498. Para verificar el valor de la media de los ingresos de la población escoge una muestra al azar de 25 ingresos. Si la media de la muestra es \$380. Determine un intervalo de confianza del 95% para la media de todos los ingresos familiares mensuales de la región. Con base en este

intervalo, ¿puede concluir que la media de los ingresos de esta población ha bajado?

Rp. σ=50\$, 380 ∓ 19.6, se rechaza que μ=400\$,

7. Un Ingeniero industrial quiere estimar la longitud media de los tornillos de alta precisión que se producen en su planta, de manera que, el estimado deberá quedar dentro de un rango de 0.04cm., con nivel de confianza 97%. El realizó un muestreo piloto y estimó la desviación estándar de la población en 0.09 cm.
   a) ¿Qué tan grande es la muestra que requiere?
   b) ¿Qué tan grande es la muestra que requiere si la población tiene tamaño 1000?

Rp. a) n =95.35≅96, b) n =87.13≅88

8. Se desea estimar la cantidad promedio mensual de dinero que pagan los 1000 alumnos de un centro de educación particular. La varianza de la población no se conoce pero se sabe que las cuentas de pago caen dentro de una amplitud de variación de 60 unidades monetarias. Halle el tamaño de la muestra necesario para estimar la media de la población con un límite para el error de estimación de 3 unidades monetarias y al nivel de confianza del 95%.

Rp. R=60, $6\sigma \cong R$, σ=60/6=10, $n=[(1.96)^2(10)^2(1000)]/[(1.96)^2(10)^2+(3)^2(999)]$=40.97≅41.

9. Con el fin de estimar el costo promedio por gastos en llamadas telefónicas en su residencia, el señor Ruiz, tomó una muestra de 36 llamadas en un mes y encontró que en promedio el tiempo de duración era de 9 minutos. El señor Ruiz sabe que la varianza del tiempo de las llamadas es 9.
   a) Halle un intervalo de confianza al 95% para la media del tiempo por llamada.
   b) Halle un intervalo de confianza al 95% para la media del costo por llamada, si el costo por minuto es de 0.10.
   c) La compañía de teléfonos ha venido considerando que en la residencia del señor Ruiz el tiempo promedio por llamada es de 11 minutos. En base a la información anterior, ¿el señor Ruiz deberá hacer un reclamo a la compañía?.

Rp. a) IC del tiempo X: 9±0.98, b) IC del costo 0.1X: 0.9±0.098, c) 11∉[8.02, 9.98], debería hacer su reclamo.

10. El ingreso mensual de cada una de las 500 microempresas de servicios de una ciudad es una variable aleatoria con media μ desconocida. La Sunat con el fin de simplificar su recaudación de estas empresas ha dispuesto que se las grave mensualmente con un 10% de sus ingresos.
    Una muestra al azar de 70 microempresas reveló que la media fue de \$710, con una desviación estándar de \$26.
    a) Estime el monto medio de los ingresos de las microempresas de la ciudad con un intervalo de confianza del 95%.

b) La Sunat se propuso lograr mensualmente una **recaudación total** de al menos \$36,000 a estas microempresas, ¿es factible que se cumplan sus metas?, ¿por qué?.

Rp. a) IC de μ: 710±5.654, b) R=0.1X, IC de μ_R:[70.4346, 71.5654], IC del total: [35,217.3, 35,782.7], 36,000 no está en el IC. No es posible

11. La empresa de distribución eléctrica piensa que mensualmente sus medidores en un distrito no registran una cantidad aleatoria X de Kwh. El distrito cuenta con 320 hogares y la empresa cobra S/. 2.3 por Kwh. A efecto de estimar el ahorro que la empresa lograría al cambiar sus medidores, la empresa ha seleccionado al azar 40 hogares del distrito, encontrando al inspeccionar sus medidores un promedio no registrado de 18 Kwh. por mes con una desviación estándar de 3 Kwh.
   a) Estime mediante un intervalo del 95% el ahorro total en soles que lograría la empresa al cambiar todos los medidores del distrito.
   b) Si la empresa quiere estimar el ahorro total al cambiar todos los medidores con un error de estimación no mayor de S/500, ¿es suficiente inspeccionar sólo 40 medidores?. Use 3Kwh como aproximación de σ..

Rp. Población finita: N=320, n=40, $\bar{x} = 18$, $\hat{s} = 3$, ES=0.444, IC al 95% de μ, 18 ∓ 0.871, del ahorro total es $320 \times 2.3(18 \mp 0.871)$. 13248 ∓ 641.056.
b) Método de población finita, n=61, no es suficiente con n=40.

12. Diariamente salen de Tarapoto hacia la costa 100 camiones cargados de diversos productos. Cada camión lleva una carga de X toneladas, cuya distribución de probabilidad se sabe que es normal. Cierto día se procedió a realizar un control, seleccionando al azar 10 de estos camiones y se procedió a pesar sus cargas, encontrándose una media de 8 toneladas con una desviación estándar de 2 toneladas.
   a) Estime mediante un intervalo de confianza del 95% la carga total transportada en un día de Tarapoto hacia la costa.
   b) ¿Qué probabilidad existe de que al realizarse este procedimiento de control durante 20 días, el verdadero valor de la carga total siempre esté, contenido en los intervalos de confianza del 95% hallados?. Asuma independencia.

Rp. a) Estd t-Student, $8 \pm 2.262 \times 2 \times (10)^{-1/2}[(100-10)/(100-1)]^{1/2}$, $100 \times [8 \pm 1.364]$, b) X: # de días en que el total cae en IC al 95%, $X \sim B(20, 0.95)$, $P[X=20]=(0.95)^{20}=0.358$.

13. Un auditor quiere estimar el monto promedio de las cuentas por cobrar de la compañía A &B. Una muestra aleatoria de 15 cuentas por cobrar escogida de un total de 400 cuentas que tiene esta compañía revela los siguientes datos en dólares:

500, 560, 560, 800, 800, 600, 600, 730,
730, 730, 640, 640, 640, 640, 870

   a) ¿Cuánto es la estimación puntual de la media de la población?

b) ¿Cuánto es el error estándar de la media?

c) ¿Qué estadística debería utilizar para estimar la media y cuál es la condición fundamental para usar esta estadística?

d) Estime la media de todas las cuentas por cobrar utilizando un intervalo de confianza del 95%. Interprete el resultado brevemente.

Rp. a) $\bar{x} = 669.333$, b) $\hat{s} = 103.886$, ES=26.348, c) Estd. t, gl=14, d) $t_0=2.145$ 669.333 ∓ 56.517

14. Una empresa de investigación quiere hacer un estudio sobre los gastos semanales de los turistas extranjeros en el Cuzco. Se sabe que la población consiste de 500 turistas extranjeros. Una muestra piloto reveló que los gastos semanales de esa población tienen distribución normal con una media de $580 y un rango de $240.

a) ¿Cuántos turistas se deberían elegir para la muestra si se quiere tener un error de estimación para la media de $20 con nivel de confianza 0.95?

b) Se escogió una muestra aleatoria de 15 turistas y ellos indicaron los siguientes gastos semanales en dólares:

200, 350, 575, 780, 890, 620, 150, 500,
700, 400, 680, 120, 300, 880, 600

Desarrolle un intervalo de confianza del 95% para la media de la población. A parir de este resultado, ¿se confirma la media de $580?.

Rp. a) 14.937, o 15, b) $\bar{x} = 516.33$, $\hat{s} = 253.92$, ES=75.747, $516. \mp 128.5$, Si,

15. En el depósito se cuentan con objetos de cierto tipo adquiridos en distintos periodos y por tanto tienen distintos costos. El comerciante estima el costo total en $30,000.00. Para verificar tal estimación se tomaron 50 costos al azar, $x_1, x_2, \ldots, x_{50}$; de tales objetos, encontrándose:,

$$\sum x_i = 900\,\$,\ \sum x_i^2 = 47{,}450\,\$^2 .$$

a) Si la estimación puntual de la media de la población es $18, ¿se puede afirmar con confianza del 97% que el error máximo de esta estimación es menor que $5?.

b) Suponga que el deposito cuenta con 2,000 objetos. Desarrolle un intervalo de confianza del 97% para el costo total. Usando este intervalo, ¿se puede aceptar como válida la estimación del comerciante?.

Rp. a) $\bar{x} = 18$, $\hat{s} = 25$, 18 ∓ 7.67, no, b) $2{,}000(18 \mp 7.57)$, 30,000 pertenece al intervalo de confianza del total, es válida la afirmación

16. Se va realizar una encuesta de muestra para estimar la media semanal de gastos en cigarrillos de los fumadores continuos quienes constituyen una población de tamaño 3000. Una encuesta a una muestra piloto revela que la desviación

estándar de la muestra es $5. El patrocinador de la encuesta quiere que el estimado esté dentro de un rango de $2.4 con un nivel de confianza del 95%. Si la encuesta tiene un gasto fijo de $2000 más un gasto de $2 por cada entrevista de la muestra. ¿Cuánto le cuesta al patrocinador este trabajo?

Rp. Población finita, n=64.2≅65, Costo=$2130

17. Se quiere estimar μ con un error máximo de estimación $e = 3$ de una población de tamaño N=1000, halle el tamaño de la muestra necesaria si se sabe que los valores de la población varía en un rango de R=100

Rp. σ≅25, n=217.56≅218

18. Halle el tamaño de muestra para estimar μ con un límite para el error de estimación de $e = 2$, si se sabe que la población está dividida en tres estratos de tamaños: $N_1 = 600$, $N_2 = 900$ y $N_3 = 1500$ y con varianzas iguales a $(25)^2$ en cada estrato.

Rp. w_1=0.2, w_2=0.3, w_3=0.5, n=333.56≅334.

Una proporción

19. La empresa de investigación estadística H&M realizó una encuesta para medir la popularidad del presidente utilizando una muestra aleatoria de 600 ciudadanos. La muestra reveló que 180 opinan a favor 360 opinan en contra y el resto de la muestra no opina al respecto.
    a) ¿Cuál es la estimación puntual de la proporción de ciudadanos en la población a favor del presidente?
    b) ¿Cuánto es el error estándar de la proporción a favor en la muestra?
    c) ¿Cuánto es el error de la estimación puntual de la proporción a favor en la población al nivel de confianza del 98%?
    d) Desarrolle un intervalo de confianza del 97% para la proporción de ciudadanos en la población a favor del presidente. Usando este intervalo, ¿es válido inferir que más del 35% de la población no está a favor del presidente?

Rp. a) 0.30, b) ES=0.0187, c) 0.04359, d) 0.30 ∓ 0.0406, No.

20. Se planea realizar una encuesta para determinar la proporción de gestantes adolescentes (menores de 16 años) utilizando una muestra aleatoria de 400 gestantes que se atienden en la maternidad de Lima. Si la muestra revela que 80 eran menores de 16 años.
    a) ¿Cuál es la proporción estimada de la población?
    b) ¿Cuáles son los límites de confianza al 99% para la proporción de adolescentes gestantes en la población?

Rp. a) 0.20, b) 0.1485 y 0.2515

21. El gerente de ventas de la tienda "CREDITOS" quiere determinar el porcentaje de clientes morosos por más de \$100. Una muestra aleatoria de 200 de tales clientes de la tienda reveló que 50 de ellos eran morosos
    a) Halle un intervalo de confianza del 98% para la proporción en la población de clientes morosos por más de \$100. ¿Es válida la estimación puntual en 33%.?
    b) Si la estimación de la proporción de la población es el intervalo [0.183, 0.317], ¿con qué grado de confianza se realizó esta estimación?.

Rp. a) 0.25 ∓ 0.07134, No, b) 97.14%.

22. El señor Lozano está pensando postular a la alcaldía de Lamas. Antes de formalizar su candidatura, decide realizar una encuesta de opinión en la localidad. Si la encuesta tiene un costo fijo de \$5000 más un costo variable de \$4 por cada entrevista, ¿cuánto le costaría este trabajo al señor Lozano, si él quiere tener un error de 2.5 puntos porcentuales con confianza del 98%?.

Rp. $n \cong (2.33)^2/4(0.025)^2 = 2171.56 \cong 2172$, costo$=5000+4\times 2172=13688$

23. Un trabajo asignado a un grupo de estudiantes consiste en realizar una encuesta para estimar la proporción de consumidores de vino nacional. Se quiere que la estimación esté dentro del 2% de la proporción de la población con un nivel de confianza del 95%. Una encuesta piloto realizada por el grupo, revela que 6 de cada 10 consumidores de vino consume vino nacional. ¿Cuál es el tamaño de muestra requerida?..

Rp. $n=(1.96)^2(0.4)(0.6)/(0.02)^2 \cong 2{,}305$.

24. Un grupo de investigación quiere estimar la proporción de estudiantes universitarios no graduados de una población de 10,000 egresados. El estimado de la proporción de la población debe estar dentro de $\mp$ 0.048. con nivel de confianza del 95%. ¿Qué tan grande es la muestra que se requiere?

Rp. $400.198 \cong 401$.

25. El gerente de comercialización de la empresa "P&C" quiere realizar una encuesta para determinar la proporción de hogares que tienen Internet. Encarga el estudio del tamaño de la muestra a un practicante de estadística. ¿Cuál de las siguientes tres opciones debería elegir el estudiante si se quiere minimizar el costo y el tiempo para realizar la encuesta?
    Estimar la proporción de la población:
    a) Con un error no mayor a 2% y con un nivel de confianza del 98%
    b) En $\mp$ 2% y con un nivel de confianza del 90%
    c) En un intervalo de amplitud 10% y con confianza del 95%?

Rp. a) $3393.06 \cong 3394$, b) $1691.27 \cong 1692$, c) $384.16 \cong 385$. Debe escoger c).

26. Dos candidatos A y B compiten como favoritos en las elecciones a Alcalde en Pamashto. Una encuesta a boca de urna reveló que el estimado de la proporción de la población a favor de A es 40% con un error de 3% al nivel de confianza del 95%, mientras que el estimado de la proporción de la población a favor de B va de 31% a 39% con nivel de confianza del 95%, ¿cuál de los dos candidatos sería el ganador absoluto?.

Rp. Cualquiera, pues hay empate técnico.

27. Un auditor toma una muestra aleatoria de 400 cuentas por cobrar y encuentra que 320 de ellas tienen deudas de al menos $700. Determine el nivel de confianza
   a) Si el porcentaje de todas las cuentas por cobrar de al menos $700 se estima de 75.76% a 84.24%.
   b) Si todas las cuentas por cobrar de al menos $700 de un total de 10,000 cuentas por cobrar se estima en el intervalo [7543, 8457].]

Rp. a) 0.966, b) 0.98.

28. Un fabricante estima en 5% la proporción de piezas defectuosas de las 5,000 producidas.
   a) Para confirmar esta estimación primero se debe escoger una muestra aleatoria. Si usted está encargado del cálculo, ¿qué tamaño de muestra recomendaría si se quiere una confianza del 95% que el error de la estimación no sea superior a 0.047?
   b) Si se escoge el mínimo tamaño de la muestra que usted recomienda y en ella se encuentran 40 piezas defectuosas, ¿se puede inferir que la estimación del fabricante es coherente con la estimación efectuada a partir de la muestra aleatoria, al nivel de confianza del 95%?.
   c) ¿Qué probabilidad existe de que al realizarse cálculos de intervalos con 20 muestras de tamaño 400 la estimación de la proporción de la población siempre esté contenido en los intervalos de confianza del 95% hallados?. Asuma independencia.

Rp. a) $n \cong 400$, b) 0.1 ∓ 0.0282, No, $p=5\%$ no pertenece al I.C., c) $(0.95)^{20}=0.358$

29. Es común usar aceros inoxidables en las plantas químicas para manejar fluidos corrosivos. Sin embargo, estos aceros tienen especial susceptibilidad al agotamiento por corrosión causada por esfuerzos en ciertos entornos. En una muestra de 295 fallas de aleaciones de acero que ocurrieron en refinerías y plantas petroquímicas durante los últimos 10 años, 118 se debieron a agrietamientos por corrosión causada por esfuerzos y fatiga de corrosión.
   a) Estime la verdadera proporción de fallas de aleaciones causadas por agrietamiento por corrosión debida a esfuerzos y fatigas de corrosión de manera que pueda atribuirle una confianza del 95% a dicha estimación.

b) Si seleccionáramos repetidamente muestras de tamaño 295 fallas de aleaciones y estableciéramos un intervalo de confianza del 95%, basado en cada uno de las muestras, ¿qué porcentaje de tales intervalos no contendrían el verdadero valor de la proporción de fallas de aleaciones causadas por esfuerzos y fatiga de corrosión?.

Rp. a) [0.344, 0.456], b) por definición el 5%

30. El instituto de estadística del gobierno quiere realizar un estudio socioeconómico de comerciantes minoristas de Lima. Una muestra piloto grande reveló que el ingreso mensual tiene una media de \$400, una desviación estándar de \$50 y que la proporción de comerciante minoristas con ingresos superiores a \$600 era de ochenta por ciento.
   a) ¿Qué tamaño de muestra de comerciantes minoristas se debe seleccionar para estimar la media con un error de \$7 y confianza 97%?
   b) ¿Qué tamaño de muestra de comerciantes minoristas se debe seleccionar para estimar la proporción de comerciantes minoristas con ingresos superiores a \$600 con un error de 5% y confianza 97%?
   c) ¿Cuántos comerciantes minoristas se deberían seleccionar para estimar la proporción de aquellos qué tienen ingresos superiores a \$600 si desconoce las estadísticas de la muestra piloto?.

Rp. a) $n \geq 240.25 \cong 241$, b) $n \geq 301.37 \cong 302$, c) $n \geq 470.89 \cong 471$.

31. Una empresa encuestadora quiere estimar el rating de una telenovela en una población que consiste de 5000 hogares con TV.
   a) ¿Qué tan grande debería ser la muestra si se desea que el estimado del porcentaje de la población que miran la telenovela tenga un error de 5% con nivel de confianza 0.95?.
   b) Desarrolle un intervalo del 95% para el porcentaje de la población que miran la telenovela si una muestra aleatoria del tamaño mínimo calculado en a) reveló que el 20% miran esa telenovela.
   c) Desarrolle un intervalo del 95% para el total de hogares en la población que miran esa telenovela.

Rp. a) $n=356.81 \cong 357$, b) 0.2 ∓ 0.04, c) 1000 ∓ 200

32. En una encuesta se entrevistó a una muestra de 150 personas de un total de 1500 para que expresen sus opiniones respecto a un proyecto local. La muestra reveló que 45 están de acuerdo, 75 están en contra y el resto de la muestra no opina. Si se infiere que entre 315 y 585 personas del total de dicha población están de acuerdo, ¿qué nivel de confianza se empleó?.

Rp. 0.9878

33. Una muestra de 100 familias escogidas al azar de una población de 1000 familias reveló el siguiente número de hijos por familia;

N# de hijos	0	1	2	3	4
N# familias	20	25	30	15	10

a) Determine un intervalo de confianza del 95% para la proporción de familias que tienen a lo más un hijo. Se puede confiar que el 50% de todas las familias tiene a lo más un hijo?

b) Halle un intervalo de confianza del 95% para el número medio de hijos por familia. Es seguro que el número medio de hijos por familia sea uno?

Rp. a) 0.45 ∓ 0.09255, Si 50% si está en el intervalo,
b) 1.7 ∓ 0.2286, No, 1 no está en el intervalo

34. Se quiere estimar p con un error máximo de estimación $e = 0.05$, hallar el tamaño de la muestra necesaria si la población es de tamaño N=2000,

Rp. p=0.5, n=333.56≅334.

Diferencia de dos medias

35. El director de presupuesto de una compañía quiere comparar el gasto de transportaciones diarias entre personal de ventas y de verificación contable. Para esto recopiló una muestra de 200 ventas y a otra muestra de 250 verificaciones contables, resultando las medias respectivas de 13 y 15 soles, y las desviaciones estándares respectivas de 3 y 4 soles. Utilizando un intervalo de confianza del 95% para la diferencia de las medias, ¿podemos concluir que la media de gastos diarios por trasportación es mayor para el personal de verificación contable?.

Rp. -2 ∓ 0.65, $\mu_1-\mu_2=0 \notin IC$ además, $\mu_1-\mu_2<0$, Si.

36. Un alto dirigente del emporio comercial Gamarra afirma que el salario promedio por semana de los hombres supera en \$13 al salario promedio de las mujeres. Una muestra aleatoria de 20 hombres y otra de 25 mujeres reveló las medias respectivas de 110 y 100 dólares. Se sabe que las dos poblaciones de salarios son normales con varianzas respectivas iguales a 100 y 64. Utilizando un intervalo con confianza del 98% para la diferencia de medias, ¿es válida la afirmación del dirigente?

Rp. 10 ∓ 5.967, $\mu_1-\mu_2=0 \notin IC$ además, $\mu_1-\mu_2=13 \in IC$, Si.

37. El jefe de personal de una empresa de confecciones quiere comparar las medias de los tiempos en minutos que operarios hombres y mujeres utilizan para confeccionar una camisa. Estudios anteriores revelan que las dos poblaciones de tiempos tienen distribución normal con varianza homogénea. Dos muestras aleatorias de tamaño 16 revelaron las siguientes estadísticas: $\bar{x}_1 = 38$, $\hat{s}_1 = 6$, y

$\bar{x}_2 = 35$, $\hat{s}_2 = 4$. Utilizando un intervalo de confianza del 95%, ¿se puede concluir que en promedio los hombres y las mujeres utilizan el mismo tiempo?

Rp. ES=1.8028, gl=30, 3 ∓ 3.68, $\mu_1-\mu_2 \in [-0.68, 6.68]$. Si.

38. Un inversionista hace un estudio con el fin de elegir una de las dos ciudades Trujillo o Piura para abrir un centro comercial. En una muestra de 21 hogares de la ciudad de Trujillo halló: $\bar{x}_1 = \$400$, $\hat{s}_1 = \$120$. En otra muestra de 16 hogares de la ciudad de Piura halló: $\bar{x}_2 = \$380$, $\hat{s}_2 = \$60$. Suponga poblaciones normales con varianzas diferentes. Usando un intervalo de confianza del 95%, ¿en cuál de las dos ciudades debería abrir la sucursal?.

Rp. ES=30.178, gl=31, 20 ∓ 61.08, $\mu_1-\mu_2=0 \in IC$ se acepta, $\mu_1=\mu_2$, en cualquiera.

39. El gerente de ventas de "REPLY" estudia el monto de los pagos con tarjeta de crédito en sus locales de Jockey Plaza y San miguel. Para realizar este trabajo, se escogieron dos muestras aleatorias de 13 y 11 días resultando los siguientes pagos en dólares con tarjeta de crédito:

Jockey Plaza: 400, 410, 420, 380, 390, 400, 410, 405, 405, 400, 410, 415, 405
San miguel: 390, 395, 380, 390, 400, 380, 370, 390, 380, 395, 390

Estos datos revelan además que las dos poblaciones de ventas tienen distribución normal con varianzas homogéneas
Halle un intervalo de confianza del 95% para la diferencia de las dos medias de las poblaciones, ¿se puede inferir que son iguales las medias de las dos poblaciones?.

Rp. Utilizando el *MCEST* se tiene: gl=22, ES=3.979, I.C: 16.573 ∓ 8.252, No.

40. Un hipermercado está estudiando la venta diaria de pollos a la brasa en dos de sus locales: Independencia y Rimac. Dos muestras aleatorias de las ventas de 13 días dieron los siguientes números de pollos vendidos:

Independencia: 12, 17, 14, 18, 09, 19, 10, 20, 15, 12, 16, 09, 14
Rimac: 12, 14, 13, 11, 12, 15, 14, 15, 11, 13, 12, 11, 14

Las muestras revelaron además que las dos poblaciones de ventas son normales con varianzas diferentes. Utilizando un intervalo de confianza del 95% para la diferencia de las dos medias poblacionales, ¿es válido inferir que las dos poblaciones tienen medias iguales?.

Rp. Utilizando el *MCEST* se tiene: gl=16, ES=1.104, I.C: 1.3846 ∓ 2.34, Si.

41. Se registraron los pesos de 15 mujeres antes de comenzar y después de terminar una dieta que tuvo una duración de 4 semanas. Los datos que se registraron en parejas, dieron las medias: $\bar{x}_1 = 70$ kg., $\bar{x}_2 = 66$ kg., y la desviación estándar de las diferencias de pesos antes y después: $\hat{s}_d = 2$ kg. Además, la población de la

diferencia de pesos tiene distribución normal. Use un intervalo de confianza del 95% para determinar si realmente la dieta baja 5 kg. promedio en 4 semanas.

Rp. $\mu_D=\mu_1-\mu_2$, gl=14, ES=0.5164, 4 ∓ 1.1077, $\mu_D=5\in$ I.C. Si.

42. La empresa de transporte de carga interprovincial "CARGO" debe decidir si compra la marca A o la marca B de neumáticos para su flota de camiones. Para esto hace un estudio de rendimiento, asignando un neumático de cada marca a las ruedas delanteras de 10 camiones y se registran en miles de kilómetros las siguientes distancias:

Camiones	1	2	3	4	5	6	7	8	9	10
Marca A	50	47	38	44	35	36	44	48	46	48
Marca B	45	43	30	39	35	31	42	44	37	46

Los resultados del experimento revelan que las diferencias de las distancias se distribuyen en forma normal.
Utilizando un intervalo de confianza del 99% para la diferencia de las dos medias, ¿se puede concluir que los promedios de rendimiento son iguales en ambas marcas?.

Rp. Utilizando *MCEST*: $\bar{d}=4.4$, $\hat{s}_d=2.716$, ES=0.859, gl=9, [2.457, 6.343], No.

Diferencia de dos Proporciones

43. La firma "PERUDIS" distribuye 2 marcas de cerveza. En una reciente encuesta se encontró que 60 de 120 prefieren la marca A y 50 de 80 prefieren la marca B. Use un intervalo de confianza del 99% para la diferencia de proporciones con el fin de determinar si son diferentes las proporciones de preferencias poblacionales de las marcas de cerveza.

Rp. -0.125 ∓ 0.18, no.

44. En octubre, 160 personas de una muestra aleatoria de tamaño 400 aprobaron la gestión de un líder político. Dos meses mas tarde, en diciembre, la mitad de otra muestra aleatoria de tamaño 500, independiente de la anterior, rechazaba tal gestión. Con un intervalo de confianza del 98%, ¿podemos concluir que dicho líder es aceptado igualmente en diciembre que en octubre?.

Rp. -0.1 ∓ 0.078, $p_1-p_2=0\notin$ IC entonces, $p_1\neq p_2$.

45. La agencia de publicidad "AVISO" realizó un estudio para comparar la efectividad de un anuncio por la radio en dos distritos. Después de difundir el aviso durante una semana, se realizó una encuesta a 900 personas seleccionadas al azar en cada uno de los distritos y se les preguntó si escucharon el aviso, resultando las proporciones 20% y 18% respectivamente. Si con estos datos se infiere que $p_1 - p_2 \in [-0.0162, 0.0562]$, ¿qué nivel de confianza se utilizó?.

Rp. $1-\alpha=0.95$

46. Dos muestras aleatorias de 250 mujeres y 200 hombres indicaron que 75 mujeres y 80 hombres consumen un nuevo producto unisex que acaba de salir al mercado. Utilizando un intervalo de confianza del 95%, ¿se puede aceptar que es igual la proporción de preferencias de mujeres y hombres en toda la población?, si no es así, ¿cuál es la relación?

Rp. $-0.1 \mp 1.96(0.0452)$, $[-0.189, -0.011]$, $p_1 \neq p_2$ además es $p_1 < p_2$.

47. En una muestra de 500 hogares de Trujillo se encontró que 50 de ellos se encuentran viendo vía satélite un programa especial de televisión. En Tarapoto, 30 hogares de una muestra aleatoria de 400 se estaban viendo el mismo programa especial. Desarrolle un intervalo de confianza del 95% para la diferencia de porcentajes reales. ¿Puede rechazarse la suposición del patrocinador de que el porcentaje de hogares que están observando el programa especial es el mismo en las dos ciudades?.

Rp. $0.025 \mp 1.96(0.0188)$, $[-0.012, 0.062]$, se acepta $p_1 = p_2$.

48. En un estudio de mercado para determinar el rating de los programas de TV del mediodía una muestra aleatoria de 400 hogares reveló que 80 estaban sintonizando el programa B de TV, 120 sintonizaban el programa G y el resto sintonizaban otra cosa. Desarrolle un intervalo de confianza del 98%. para la diferencia de proporciones. ¿Es la proporción global de televidentes que sintonizan el programa B igual al que sintonizan G?. Si no es así, ¿cuál es la relación?

Rp. -0.1 ∓ 2.33 (0.0304), $[-0.171, -0.029]$, $p_1 \neq p_2$ además es $p_1 < p_2$.

Varianzas

49. En una muestra aleatoria de 13 tiendas se encontró que las ventas de la semana de un determinado producto de consumo popular tiene una desviación estándar $\hat{s} = \$6$. Estudios anteriores revelan que las ventas del producto tienen distribución normal. Estime la varianza poblacional mediante un intervalo de confianza del 95%.

Rp. $(432/23.34) \leq \sigma^2 \leq (432/4.4)$, entonces, $18.51 \leq \sigma^2 \leq 98.18$.

50. Una muestra aleatoria de 16 sobres de cierto producto cuyos pesos se distribuyen normalmente ha dado una desviación estándar de 0.6 gramos.
   a) Halle un intervalo de confianza bilateral del 95% para la desviación estándar. ¿Es válido inferir que la desviación estándar de los pesos de tales sobres es 0.25?
   b) Para un intervalo unilateral del 95% para la desviación estándar, ¿qué tan grande puede ser la desviación estándar de los pesos?

Rp. a) $0.1964 \leq \sigma^2 \leq 0.8626$, $0.4432 \leq \sigma \leq 0.9288$, No, b) $\sigma^2 \leq 0.7438$, $\sigma \leq 0.8624$.

51. Una gran corporación que realiza ventas de productos de consumo masivo decidió analizar la dispersión de las ventas semanales de un producto determinado en sus 400 tiendas. Tales ventas se distribuyen aproximadamente normal. Si en una muestra aleatoria de 15 de sus tiendas se encontró las siguientes ventas semanales en dólares:

700, 739, 695, 710, 724, 715, 720, 723, 700, 750, 695, 760, 689, 735, 670

Estime la desviación estándar de las ventas del producto mediante un intervalo de confianza del 95%.

Rp. $\hat{s} = 24.454$, $320.5196 \leq \sigma^2 \leq 1487.028$, $17.9 \leq \sigma \leq 38.562$

52. Una de las maneras de medir el grado de satisfacción de los empleados de una misma categoría en cuanto a la política salarial es a través de las varianza de sus salarios. La fábrica A afirma ser más homogénea en la política salarial que la fábrica B. Para verificar esa afirmación, se escogieron una muestra aleatoria de 10 salarios de A y otra de 13 salarios de B, obteniendo las dispersiones $\hat{s}_A = 50$, $\hat{s}_B = 30$. Registros anteriores indican que los salarios de A y de B tienen distribuciones normales. ¿Cuál sería su conclusión si utiliza un intervalo del 95% para el cociente de las dos varianzas?.

Rp. $0.80 \leq \sigma_1^2 / \sigma_2^2 \leq 10.76$, . Si.

53. Utilice un paquete de computo estadístico (por ejemplo el *MCEST*) para verificar por el método de intervalo del cociente de dos varianzas con nivel de confianza de 0.95 que:

a) son iguales las dos varianzas poblacionales en el ejercicio 39
b) son diferentes las dos varianzas poblacionales en el ejercicio 40

Rp. a) $0.328 \leq \sigma_1^2 / \sigma_2^2 \leq 5.87$, Si., b) $1.411 \leq \sigma_1^2 / \sigma_2^2 \leq 27.66$, . Si.

54 *. **Use un paquete de computo estadístico para resolver el siguiente problema:**

Con referencia a la hoja de cálculo del *estudio socioeconómico de universitarios de Lima* (ver apéndice)

a) Determine un intervalo de confianza de 95%, para el gasto medio mensual en educación por familia.
b) Determine un intervalo de confianza de 95%, para la proporción de familias que poseen casa propia.

Rp. a) 1835.7576 ∓ 354.09595, b) 0.606 ∓ 0.1667.

Capítulo 10

PRUEBAS DE HIPOTESIS

Introducción.

El objetivo es dar algunos métodos que se usan para tomar decisiones sobre poblaciones, a partir de los resultados de una muestra aleatoria escogida de esa población. Para llegar a **tomar *decisiones estadísticas* se debe partir de afirmaciones o conjeturas con respecto a la población** en el que estamos interesados. Tales suposiciones, pueden ser verdaderas o no. Una conjetura hecha sobre una población o sobre sus parámetros deberá ser sometida a comprobación experimental con el propósito de saber si los resultados de una muestra aleatoria extraída de esa población, contradicen o no tal conjetura.

10.1 Hipótesis estadísticas.

Definición. Se denomina **hipótesis estadística** a cualquier afirmación o conjetura que se hace acerca de la distribución de una o más poblaciones.

La afirmación o conjetura puede referirse bien a la forma o tipo de distribución de probabilidad de la población o bien referirse al valor o valores de uno o más parámetros de la distribución conocida su forma.

En las aplicaciones básicas, se asume dada la forma de la distribución de la población. En este caso, las hipótesis estadísticas consisten en suponer que los parámetros, que definen a la población, toman determinados valores numéricos.

Por **ejemplo**, son hipótesis estadísticas:

1. La longitud media de un tipo de objetos es 10 centímetros.
2. La proporción de objetos defectuosos producidos por cierto proceso nunca es superior al 8%.
3. La varianza de la longitud de cierto tipo de objetos es 0.25 cm^2..
4. Son iguales las medias de dos tipos de mediciones independientes X e Y que se distribuyen normalmente con varianza común σ^2.

Hipótesis simple y compuesta.

Definición. Se denomina **hipótesis simple** a cualquier hipótesis estadística que especifica completamente la distribución de la población, es decir, especifica la forma de la distribución y el valor de su(s) parámetro(s).

Si una hipótesis no especifica completamente la distribución de la población se dice que es una **hipótesis compuesta**.

Por **ejemplo**, la hipótesis que establece que el ingreso mensual promedio de los empleados de cierta ciudad es $\mu = \$500$, suponiendo que los ingresos mensuales se distribuyen según la normal con desviación estándar conocida $\sigma = \$30$; es una hipótesis simple, pues, especifica completamente la distribución de la población. En cambio, si se supone que los ingresos mensuales se distribuyen según la normal con desviación estándar conocida $\sigma = \$30$ y se afirma que el ingreso promedio mensual es $\mu \neq 500$ ó $\mu < 500$ ó $\mu > 500$, entonces, la hipótesis referente a la media es una hipótesis compuesta, pues, no especifica la media de la distribución de la población de los ingresos.

Hipótesis nula y alternativa

Definición. Se denomina **hipótesis nula** y se representa por H_0 a la hipótesis que es aceptada provisionalmente como verdadera y cuya validez será sometida a comprobación experimental. Los resultados experimentales nos permitirán seguir aceptándola como verdadera o si, por el contrario, debemos rechazarla como tal.

Toda hipótesis nula va acompañada de otra hipótesis alternativa.

Definición. Se denomina **hipótesis alternativa** y se representa por H_1 o por H_A a la suposición contraria a la hipótesis nula. La hipótesis alternativa H_1 se acepta en caso de que la hipótesis nula H_0 sea rechazada.

NOTA. La hipótesis nula siempre debe contener el signo igual por que es la hipótesis que se va a probar y es necesario que incluya un valor específico del parámetro.

Por **ejemplo**, en general, si se asume que θ_0 es un valor del parámetro desconocido θ (donde θ representa a μ, p, σ^2, etc) de una población cuya distribución se supone conocida, entonces, son hipótesis nulas y alternativas respectivamente las siguientes afirmaciones:

1) $H_0: \theta = \theta_0$, y $H_1: \theta \neq \theta_0$
2) $H_0: \theta \leq \theta_0$, y $H_1: \theta > \theta_0$
3) $H_0: \theta \geq \theta_0$, y $H_1: \theta < \theta_0$

Prueba de una hipótesis estadística

Para tomar decisiones estadísticas, se requieren de las dos hipótesis: la *hipótesis nula* y la *hipótesis alternativa* referidas a un parámetro θ.

La **prueba de una hipótesis estadística** es un proceso que nos conduce a tomar la decisión de aceptar o rechazar la hipótesis nula H_0 en contraposición de la hipótesis alternativa H_1 y con base en los resultados de una muestra aleatoria seleccionada de la población en estudio.

La hipótesis nula H_0 es la primera hipótesis que se plantea, y debe ser establecida de manera que *especifique un valor* θ_0 del parámetro θ en estudio. Por esta razón, algunos autores plantean la hipótesis nula $H_0 : \theta = \theta_0$ aun para los casos 2) y 3) del ejemplo anterior.

La aceptación de la hipótesis nula significa que los datos de la muestra no proporcionan evidencia suficiente para refutarla. El rechazo significa que los datos de la muestra proporcionan evidencia suficiente de que la hipótesis nula es falsa.

Tipos de pruebas de hipótesis.

El **tipo de prueba** depende básicamente de la hipótesis alternativa H_1

Se denomina **prueba de una cola** a toda prueba de hipótesis donde la alternativa H_1 es unilateral.

Si la alternativa es bilateral, la prueba se denomina **prueba de dos colas.**

Por **ejemplo**, si se asume que θ_0 es un valor del parámetro desconocido θ:

1) La prueba de hipótesis $H_0 : \theta = \theta_0$ contra $H_1 : \theta \neq \theta_0$ se denomina **prueba bilateral o de dos colas.**
2) La prueba de hipótesis $H_0 : \theta = \theta_0$ contra $H_1 : \theta > \theta_0$ se denomina **prueba unilateral de cola a la derecha.**
3) La prueba de hipótesis $H_0 : \theta = \theta_0$ contra $H_1 : \theta < \theta_0$ se denomina **prueba unilateral de cola a la izquierda.**

Errores tipo I y tipo II, y Nivel de significación

Al tomar la decisión de *aceptar* o *rechazar* la hipótesis nula $H_0 : \theta = \theta_0$ con base en los resultados obtenidos de una muestra aleatoria seleccionada de la población en estudio; **hay cuatro posibles situaciones que determinan si la decisión tomada es correcta o incorrecta**, como se muestra en la tabla 10-1.

Definición. Se denomina **error tipo I,** al error que se comete al rechazar la hipótesis nula H_0 cuando es verdadera.

Definición. Se denomina **error tipo II,** al error que se comete al aceptar la hipótesis nula H_0 cuando es falsa.

La probabilidad de cometer un error tipo I se denota por α. Es decir,

$$\alpha = P[\text{error tipo I}] = P[\text{rechazar } H_0 \text{ cuando } H_0 \text{ es verdadera}].$$

La probabilidad de la **decisión correcta** de aceptar H_0 cuando es verdadera es $1-\alpha$

La probabilidad de cometer un error tipo II se denota por β. Es decir,

$$\beta = P[\text{error tipo II}] = P[\text{aceptar } H_0 \text{ cuando } H_0 \text{ es falsa}].$$

La probabilidad de la **decisión correcta** de rechazar H_0 cuando es falsa es $1-\beta$.

Tabla 10-1

Decisión	H_0 verdadera	H_0 falsa
Rechazar H_0	Error tipo I Probab: α	Decisión correcta Probab: $1-\beta$
Aceptar H_0	Decisión correcta Probab: $1-\alpha$	Error tipo II Probab: β

Definición. Se denomina **nivel de significación** de una prueba de hipótesis a la probabilidad de cometer un error de tipo I.

El nivel de significación también conocido como nivel de riesgo, se fija previamente por lo general en $\alpha = 0.05$ o $\alpha = 0.01$.

Si para un valor dado de α, se rechaza la hipótesis H_0, entonces se dice que los resultados muestrales obtenidos, no sólo son diferentes por efectos del azar, si no que son realmente **significativamente diferentes** al nivel $\alpha \times 100\%$, es decir; se espera que de 100 resultados muestrales en $\alpha \times 100\%$ de las veces se rechazará la hipótesis nula H_0 cuando realmente es verdadera.

Definición. La **potencia** de una prueba es la probabilidad de tomar la decisión acertada de, rechazar H_0 cuando ésta es falsa o de aceptar H_1 cuando ésta es verdadera La potencia de una prueba es pues $1-\beta$.

Para una muestra aleatoria de tamaño n seleccionada de la población en estudio, si α aumenta, entonces β disminuye, y si β aumenta, entonces α disminuye. Por supuesto, en todo proceso de toma de decisiones sobre hipótesis estadísticas, es deseable disminuir las probabilidades de cometer esos dos tipos de errores. Este problema es tratado ampliamente en otros textos de estadística.

Región crítica y regla de decisión

Después de plantear la hipótesis nula H_0 y su correspondiente alternativa H_1 referentes a un parámetro θ y especificado el tamaño α del nivel de significación de la prueba de H_0 contra H_1, se deberá **determinar una estadística** Θ correspondiente al parámetro cuya distribución muestral se conozca. Por ejemplo, si las hipótesis H_0 y H_1 se expresan en términos de la media poblacional μ, entonces, se seleccionará la media muestral $\overline{X}$ como la estadística apropiada para efectuar la prueba.

Si se supone que la hipótesis nula $H_0: \theta = \theta_0$ es verdadera; entonces, la distribución de probabilidad de la estadística Θ queda bien definida por esta hipótesis, ya que esta hipótesis especifica completamente la distribución.

En la distribución de probabilidad de la estadística fijada por la hipótesis nula $H_0: \theta = \theta_0$ se establece la ***regla de decisión*** de acuerdo con la cual se rechazará o por el contrario se aceptará la hipótesis H_0. El rechazo de la hipótesis nula H_0 implica la aceptación de H_1.

La regla de decisión implica la división de la distribución muestral de la estadística Θ de la prueba en dos partes mutuamente excluyentes: la ***región de rechazo*** o ***región crítica*** *(R.C.)* de H_0, y la ***región de aceptación*** *(R.A.)* o *no rechazo* de H_0. Esta división depende de la hipótesis alternativa H_1, del nivel de significación α y de la distribución muestral de la estadística.

Por **ejemplo**, supongamos que se tiene una población normal $N(\mu, 9)$ con **varianza conocida** $\sigma^2 = 9$ y que se trata de **probar** o ***docimar*** la hipótesis nula $H_0: \mu = 70$ contra $H_1: \mu > 70$.

Dado que $\overline{X}$ es un buen estimador de μ utilizaremos esta estadística para determinar la región crítica y la regla de decisión de esta prueba. Como **estamos interesados en la discriminación entre** $\mu = 70$ **y valores de** $\mu > 70$ **parece razonable que debamos rechazar** H_0 **si** $\overline{X} - 70$ **es muy grande, esto es si** $\overline{X} > K$, donde K **es un *valor crítico* de la prueba** que vamos a determinar.

Si **se supone verdadera** la hipótesis $H_0: \mu = 70$, entonces, la distribución de la media $\overline{X}$ es normal con media $\mu = 70$ y desviación estándar $\sigma = 3$.

En consecuencia la distribución de

$$Z = \frac{\overline{X} - 70}{3/\sqrt{n}}$$

es normal $N(0,1)$.

Para una muestra aleatoria de tamaño $n = 40$ y la probabilidad de error tipo I, $\alpha = 0.5$ se tiene (ver figura 10.1).

$$0.05 = P[\text{rechazar } H_0 \text{ cuando } H_0 \text{ es veradera}] = P[\overline{X} > K/\mu = 70]$$

$$0.05 = P\left[\frac{\overline{X} - 70}{3/\sqrt{40}} > \frac{K - 70}{3/\sqrt{40}}\right] = P\left[Z > \frac{K - 70}{0.474}\right].$$

De la tabla normal estándar $N(0,1)$ se obtiene:

$$\frac{K - 70}{0.474} = 1.645, \text{ luego } K = 70 + 1.645 \times 0.474 = 70.78\,.$$

Por tanto, la *región crítica* en el rango de variación de $\overline{X}$ es el intervalo

$$R.C. =]70.78, +\infty[\,.$$

La ***regla de decisión*** es: si $\overline{x}$ es el valor de $\overline{X}$ obtenido a partir de una muestra aleatoria de tamaño 40, se rechazará H_0 si $\overline{x} > 70.78$.

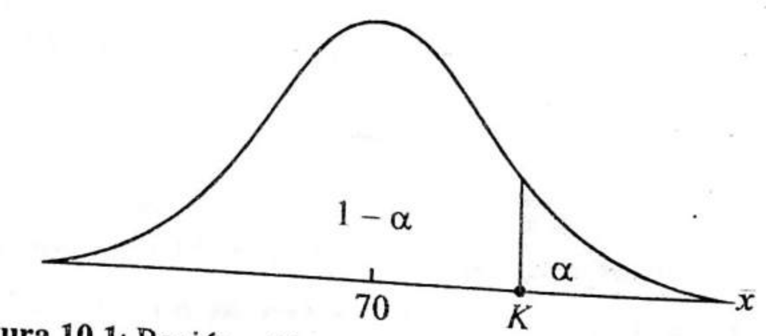


Figura 10.1: Región crítica cola a la derecha en la variable $\overline{X}$

Procedimiento de la prueba de hipótesis

Previamente debe formularse el problema estadístico, determinar la variable en estudio y el método estadístico adecuado para la solución del problema. El procedimiento general de la prueba de una hipótesis de parámetro general θ se resume en los siguientes pasos:

1) Formular la *hipótesis* nula $H_0 : \theta = \theta_0$ y la hipótesis alternativa adecuada $H_1 : \theta \neq \theta_0$ ó $H_1 : \theta > \theta_0$ ó $H_1 : \theta < \theta_0$
2) Especificar el tamaño α del *nivel de significación*.
3) Seleccionar la *estadística* apropiada a usar en la prueba.

4) Establecer la *regla de decisión*, determinando la región crítica de la prueba.
5) *Calcular* el valor del estadístico de la prueba a partir de los datos de la muestra.
6) Tomar la *decisión* de rechazar la hipótesis H_0 si el valor de la estadística de la prueba está en la región crítica. En caso contrario, no rechazar H_0.

10.2 Pruebas de hipótesis acerca de la media μ: Varianza σ^2 supuesta conocida

Sea $\overline{X}$ la media de una muestra alcatoria de tamaño n seleccionada de una población con media μ y varianza σ^2 supuestamente conocida.

Si la ***población es normal*** $N(\mu, \sigma^2)$, entonces, la distribución de la estadística $\overline{X}$ es exactamente normal $N(\mu, \sigma^2/n)$ para cualquier valor de n $(n \geq 2)$. Si la ***población no es normal***, pero el tamaño de la muestra es suficientemente grande $(n \geq 30)$, entonces, la distribución de $\overline{X}$ es aproximadamente normal $N(\mu, \sigma^2/n)$. Entonces,

La estadística para la prueba acerca de μ con varianza σ^2 conocida es

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución es exacta o aproximadamente normal estándar $N(0,1)$, según sea la población normal o no

Si se supone verdadera la hipótesis nula: $H_0 : \mu = \mu_0$, la estadística especificada por esta hipótesis es entonces:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, dado el nivel de significación α, en la distribución de $Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n})$, que es normal $N(0,1)$, se determina el valor crítico $z_{1-\alpha/2}$ tal que la probabilidad de rechazar H_0 cuando se supone verdadera sea (ver figura 10.2)

$$P[Z < -z_{1-\alpha/2}] = \alpha/2 \text{ o } P[Z > z_{1-\alpha/2}] = \alpha/2.$$

Luego, ***la región crítica o de rechazo H₀ en el rango de variación de Z*** es:

$$R.C. = \{Z < -z_{1-\alpha/2} \text{ o } Z > z_{1-\alpha/2}\}.$$

Por otro lado, la probabilidad de aceptar H_0 cuando se supone verdadera es:

$$P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1 - \alpha.$$

Resultando la ***región de aceptación de H₀***: $R.A. = \{-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\}$.

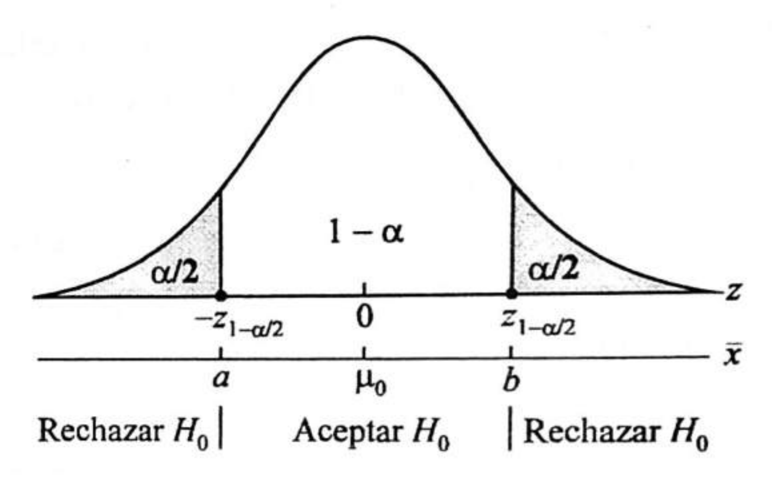


Figura 10.2: Región crítica bilateral en escalas z y $\overline{x}$

Regla de decisión es: Si $z_k = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$ es un valor de Z obtenido de la muestra, entonces, se rechazará H_0 con riesgo igual a α, si $\overline{x} \in R.C.$ (o si $\overline{x} \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.2).

Si se rechaza H_0 se dice que el *valor* z_k *es significativo* con un riesgo cuyo valor es α.

NOTA. (Región crítica en $\overline{X}$)

Si se sustituye $Z = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ en RC resulta ***la región crítica o de rechazo H₀ en el rango de variación de*** $\overline{X}$:

$$R.C. = \{\overline{X} < a \text{ o } \overline{X} > b\}.$$

donde $a = \mu_0 - z_{1-\alpha/2}(\sigma/\sqrt{n})$, y $b = \mu_0 + z_{1-\alpha/2}(\sigma/\sqrt{n})$

La región de aceptación *de* H_0 en $\overline{X}$ es el intervalo: $R.A. = [a \leq \overline{X} \leq b]$.

La *regla de decisión* es: Si $\bar{x}$ es el valor de $\bar{X}$ obtenido a partir de una muestra aleatoria, se rechazará H_0 con un riesgo α, si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.2).

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu > \mu_0$, dado el nivel de significación α, en la distribución de $Z = (\bar{X} - \mu_0)/(\sigma / \sqrt{n})$ que es normal $N(0,1)$, se determina el valor $z_{1-\alpha}$ tal que (figura 10.3),

$$P[Z > z_{1-\alpha} \,/\, H : \mu = \mu_0 \text{ verdadera}] = \alpha$$

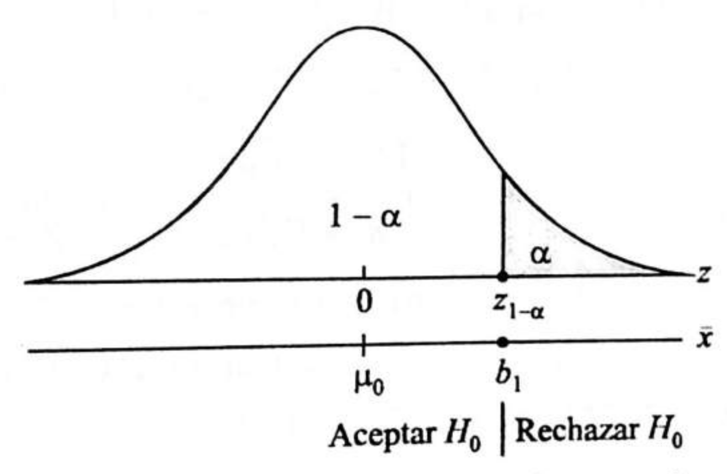


Figura 10.3: Región crítica cola a la derecha en escalas z y $\bar{x}$

Luego, *la región crítica* ***o de rechazo H₀*** *en el rango de variación de* Z es:

$$R.C. = \{Z > z_{1-\alpha}\}$$

La *región de aceptación de* H_0 es el intervalo: $R.A. = \{Z \leq z_{1-\alpha}\}$.

La *regla de decisión* es: Si $z_k = (\bar{x} - \mu_0)/(\sigma / \sqrt{n})$ es un valor de Z obtenido a partir de una muestra, se rechazará H_0 si $z_k \in R.C.$ (o si $z_k \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.3).

NOTA. (Región crítica en $\bar{X}$)

Si se sustituye $Z = (\bar{X} - \mu_0)/(\sigma / \sqrt{n})$ en RC resulta la región crítica ***o de rechazo H₀*** en el rango de variación de $\bar{X}$: $R.C. = \{\bar{X} > b_1\}$, donde:

$b_1 = \mu_0 + z_{1-\alpha}(\sigma/\sqrt{n})$

La *región* de aceptación *de* H_0 es el intervalo: $R.A. = \{\overline{X} \leq b_1\}$

La *regla de decisión* es: Siendo $\overline{x}$ el valor de $\overline{X}$ obtenido a partir de una muestra aleatoria de tamaño *n*, se rechazará H_0 con un riesgo α, si $\overline{x} \in R.C.$ (o si $\overline{x} \notin R.A.$). No se rechazará H_0 en caso contrario.(Figura 10.3).

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu < \mu_0$, dado el nivel de significación α, en la distribución de $Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n})$, se puede determinar el valor crítico $z_{1-\alpha}$ tal que (ver figura 10.4)

$$P[Z < -z_{1-\alpha} / H : \mu = \mu_0 \text{ verdadera}] = \alpha$$

Luego, *la región crítica* ***o de rechazo H₀*** *en el rango de variación de* Z es:

$$R.C. = \{Z < -z_{1-\alpha}\}$$

En consecuencia, la *región de aceptación de* H_0 es: $R.A. = \{Z \geq -z_{1-\alpha}\}$.

Regla de decisión: Si $z_k = (\overline{x} - \mu_0)/(\sigma / \sqrt{n})$ es un valor de Z obtenido a partir de una muestra, se rechazará H_0 con un riesgo, α; si $\overline{x} \in R.C.$ (o si $\overline{x} \notin R.A.$). No se rechazará H_0 en caso contrario (figura 10.4).

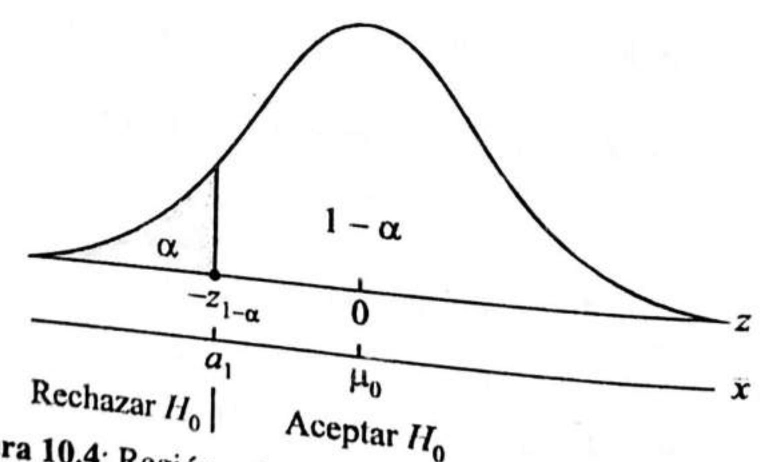


Figura 10.4: Región crítica cola a la izquierda en escalas z y $\overline{x}$

NOTA. (Región crítica en $\overline{X}$)

Si se sustituye $Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n})$ en RC y manipulando algebraicamente, se obtiene la región crítica ***o de rechazo H₀*** en el rango de variación de $\overline{X}$: $RC = \{\overline{X} < a_1\}$, donde: $a_1 = \mu_0 - z_{1-\alpha}(\sigma/\sqrt{n})$

Región de aceptación de H_0: $RA = \{\overline{X} \geq a_1\}$.

Regla de decisión es: Si $\bar{x}$ es un valor de $\bar{X}$ obtenido a partir de una muestra aleatoria de tamaño n, se rechazará H_0 con un riesgo α si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$). No se rechazará H_0 en caso contrario (figura 10.4).

EJEMPLO 10.1.

El gerente de ventas de la empresa H&B afirma que las venta diarias se distribuyen según el modelo de la probabilidad normal con una media de \$400 y una desviación estándar de \$20.

Para verificar la hipótesis con respecto a la media, un analista escogió una muestra aleatoria de las ventas de 100 días y encontró que la media de las ventas es igual a \$395. Si el analista utiliza una hipótesis alternativa bilateral y el nivel de significación del 5%, ¿cuál sería su conclusión?.

SOLUCION.

Sea X la variable aleatoria que denota las ventas diarias de la empresa. Se supone que la distribución de X es $N(\mu, (20)^2)$.

1. *Hipótesis*: $H_0 : \mu = 400$ contra $H_1 : \mu \neq 400$
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Población normal con varianza conocida, la estadística es

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución es normal $N(0,1)$.

4. *Región crítica*: Si se supone verdadera la hipótesis nula $H_0 : \mu = 400$ para $\alpha = 0.5$ y la alternativa bilateral, en la distribución de:

$$Z = (\bar{X} - 400)\big/(20/\sqrt{100})$$

-1,96 +1,96

se encuentra el valor crítico: $z_{1-\alpha/2} = z_{0.975} = 1.96$

Luego, la región crítica de la prueba en la variable Z es el conjunto

$$RC = \{Z < -1.96 \ \text{ o } \ Z > 1.96\}$$

5. *Cálculos*: De los datos se tiene: $n = 100$, $\bar{x} = 395$, $\sigma = 20$,

$$z_k = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{395 - 400}{2} = -2.5.$$

6. *Decisión*: Dado que $z_k = -2.5 \in R.C$, el analista debería rechazar la hipótesis del gerente con respecto a la media de ventas de la empresa con un riesgo de cometer error tipo I igual a 0.05.

NOTA. En el rango de variación de $\overline{X}$ la región crítica de la prueba es:

$$R.C. = \{\overline{X} < 400 - 1.96 \times 2 \quad \text{o} \quad \overline{X} > 400 + 1.96 \times 2\}$$

$$R.C. = \{\overline{X} < 396.08 \quad \text{o} \quad \overline{X} > 403.92\}$$

El hecho que $\bar{x} = 395 \in R.C.$, el analista debería rechazar H_0. Esto es, la media de la muestra difiere de la media supuesta de la población en forma significativa al 5%.

NOTA. (Regla de decisión en Intervalo de confianza)

La **prueba bilateral** de la hipótesis nula $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$ a un nivel de significación dado α, **equivale a calcular el intervalo de confianza** ($I.C.$) de $(1-\alpha)\times 100\%$ para el parámetro μ y luego rechazar la hipótesis nula $H_0 : \mu = \mu_0$ si es que $\mu_0 \notin I.C.$

En efecto, si $\bar{x}$ es un valor de $\overline{X}$ no se rechazará $H_0 : \mu = \mu_0$ si el valor

$$z_k \in R.A. = [-z_{1-\alpha/2}, z_{1-\alpha/2}], \text{ donde } z_k = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$$

o, si

$$-z_{1-\alpha/2} \leq \frac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}$$

Esto es, no se rechazará $H_0 : \mu = \mu_0$ si

$$\bar{x} \in R.A. = [\mu_0 - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

o equivalentemente si μ_0 se encuentra dentro del intervalo de confianza ($I.C.$) del $(1-\alpha)\times 100\%$ para μ:

$$\mu_0 \in I.C. = [\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

Por tanto, se rechazará H_0 con riesgo α si,

$$\bar{x} \notin R.A. \quad \text{o si} \quad \mu_0 \notin I.C$$

Por **ejemplo**, continuando con el ejemplo 10.1, para $\alpha = 0.05$ se tiene:

$$I.C. = \left[\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right] = [391.08, 398.92]$$

$$R.A. = \left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [396.08, 403.92].$$

Dado que $\mu_0 = 400 \notin I.C.$ (o que $\bar{x} = 395 \notin R.A.$) se debería rechazar H_0 con un riesgo del 5%.

NOTA. (Método del valor de la probabilidad *P*)

Otra forma de establecer la regla de decisión, en estadística aplicada, es calculando el valor P, a partir del *valor absoluto* de $z_k = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$, (que se obtiene de la muestra), de manera que

a) $P = P[Z < -z_k] + P[Z > z_k] = 2P[Z > z_k]$ (para una prueba bilateral).

b) $P = P[Z > z_k]$ (para una prueba unilateral cola a la derecha).

c) $P = P[Z < -z_k]$ (para una prueba unilateral cola a la izquierda).

Si el valor de $P < \alpha$, entonces, se rechazará H_0. No se rechazará H_0, en caso contrario.

Las pruebas de hipótesis con el paquete estadístico *MCEST* contienen el método del valor *P*.

En el **ejemplo 10.1**, el valor absoluto de z_k es igual a 2.5, entonces,

$$P = 2P[Z > z_k] = 2P[Z > 2.5] = 2(0.0062) = 0.0124.$$

Dado que $P = 0.0124 < \alpha = 0.05$, se debe rechazar H_0, con un riesgo $\alpha = 0.05$. Observe que z_k es significativo en un valor mucho menor que $\alpha = 0.05$ y que este valor de z_k sólo ocurrirá en 124 casos de 10,000 experimentos.

Una región crítica de tamaño 0.0124 es muy pequeña y, por lo tanto, es poco probable que se cometa error tipo I.

NOTA. Si la población no tiene distribución normal y si la varianza es desconocida, para probar hipótesis acerca de la media μ se utiliza la estadística:

$$Z = \frac{\bar{X} - \mu_0}{\hat{s}/\sqrt{n}}, \text{ que es aproximadamente normal } N(0,1)$$

sólo si, el tamaño de la muestra es grande. Esto es, si $n \geq 30$.

En este caso, la desviación estándar σ se estima puntualmente por $\hat{s}$ de la muestra.

EJEMPLO 10.2.

Al estudiar si conviene tener o no una sucursal en la ciudad de Tarapoto, la gerencia de una gran tienda comercial de Lima, establece el siguiente criterio para tomar una decisión: Abrir la sucursal sólo si el ingreso promedio familiar mensual en dicha ciudad es no menos de \$500 y no abrirla en caso contrario. Si una muestra aleatoria de 100 ingresos familiares de esa ciudad ha dado una media de \$480 y una desviación estándar de \$80.

a) ¿Cuál es la decisión a tomar al nivel de significación del 5%?.

b) ¿Con qué probabilidad esta prueba detecta la diferencia igual a \$30 en el promedio de ingresos y por debajo de lo que se indica en la hipótesis nula?.

c) Calcule la potencia de la prueba si el ingreso promedio realmente es \$464.

d) Calcule el valor de P y de su comentario sobre este valor.

SOLUCION.

Sea X la variable aleatoria que representa los ingresos familiares mensuales de los pobladores de Tarapoto.

a) 1. *Hipótesis*: $H_0 : \mu \geq 500$) (se abre la sucursal).

$H_1 : \mu < 500$ (no se abre la sucursal).

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Población no normal, $n = 100$, $\sigma = \hat{s}$ =80, por teorema central del límite la estadística apropiada es:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución es aproximadamente normal $N(0,1)$.

4. *Región crítica*: Si se supone verdadera la hipótesis nula $H_0 : \mu = 500$, para $\alpha = 0.5$ y la alternativa unilateral cola a la izquierda, en la distribución de $Z = (\overline{X} - 500)/(80/\sqrt{100})$, se encuentra el valor crítico

$$z_\alpha = z_{0.05} = -1.645\,.$$

Luego, la región crítica o de rechazo de H_0 en Z es: $RC = \{Z < -1.645\}$

5. *Cálculos*: De la muestra se tiene, $\bar{x} = 480$

$$z_k = \frac{\bar{x} - 500}{80/\sqrt{100}} = \frac{480 - 500}{8} = -2.5,$$

6. *Decisión*: Dado que $z_k = -2.5 \in R.C.$. Deberíamos rechazar H_0 y decidir no abrir la sucursal en la ciudad de Tarapoto. Con un riesgo de 0.05.

b) La región crítica o de rechazo de H_0 en la variable de la media $\overline{X}$ es:

$$RC = \{Z < -1.645\} = \{\overline{X} < 500 - 1.645 \times 8\} = \{\overline{X} < 486.84\}$$

La prueba anterior detecta la diferencia igual a \$30 en el promedio de ingresos por debajo de lo que se indica en la hipótesis nula si se rechaza H_0 cuando $\mu = 470$. Entonces, la probabilidad de la decisión correcta de rechazar H_0 cuando es falsa es:

$$P[\overline{X} < 486.84 / \mu = 470] = P\left[Z < \frac{486.84 - 470}{80/\sqrt{100}}\right] = P[Z < 2.11] = 0.9826$$

c) La potencia de la prueba es: $1 - \beta = 1 - 0.0021 = 0.9979$, donde la probabilidad β de aceptar H_0 cuando realmente es $\mu = 464$ (error tipo II) es:

$$\beta = P[\overline{X} \geq 486.84 / \mu = 464] = P\left[Z \geq \frac{486.84 - 464}{80/\sqrt{100}}\right] = P[Z \geq 2.86]$$

$$\beta = P[Z \geq 2.86] = 0.0021.$$

d) En la distribución de Z se halla: $P = P[Z < -2.5] = 0.0062$. El valor z_k **es significativo** en un valor menor de 0.0062. Este valor de z_k sólo ocurrirá en 62 casos de 10,000 experimentos.

EJEMPLO 10.3. (Tamaño de la muestra)

Suponga que X es una población normal con media μ (desconocida) y con varianza σ^2 supuesta conocida.

Dadas las probabilidades α y β de cometer errores tipo I y tipo II respectivamente, determinar el tamaño n de la muestra requerida para probar las hipótesis simples

$$H_0 : \mu = \mu_0 \text{ contra } H_1 : \mu = \mu_1 \text{ donde } \mu_1 < \mu_0.$$

SOLUCION.

Sea K el punto crítico en la variable $\overline{X}$ de la prueba unilateral cola a la izquierda de H_0 contra H_1 (figura 10.5), entonces,

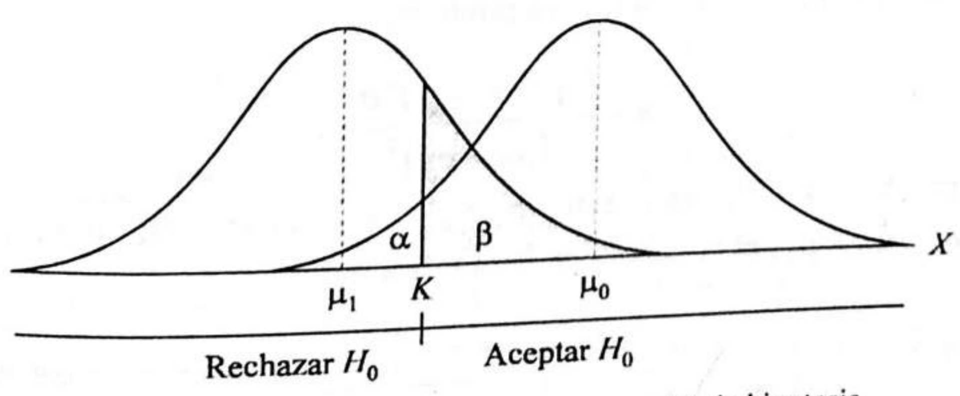


Figura 10.5: Tamaño de muestra en prueba de hipotesis

$$\alpha = P[\text{error tipo I}] = P[\text{rechazar } H_0 \;/\; H_0 : \mu = \mu_0 \text{ es verdadera}]$$

$$\alpha = P\left[\overline{X} < K / \mu = \mu_0\right] = P\left[\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{K - \mu_0}{\sigma/\sqrt{n}}\right] = P\left[Z < \frac{K - \mu_0}{\sigma/\sqrt{n}}\right]$$

de donde resulta (por ser α menos del 50% del área total y $z_\alpha = -z_{1-\alpha}$)

$$\frac{K - \mu_0}{\sigma/\sqrt{n}} = -z_{1-\alpha}, \qquad K = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \tag{1}$$

De igual manera,

$$\beta = P[\text{error tipo II}] = P[\text{aceptar } H_0 \;/\; H_0 : \mu = \mu_0 \text{ es falsa}]$$

$$\beta = P[\text{aceptar } H_0 \;/\; H_1 : \mu = \mu_1 \text{ es verdadera}]$$

$$\beta = P[\overline{X} \geq K / \mu = \mu_1] = P\left[\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} \geq \frac{K - \mu_1}{\sigma/\sqrt{n}}\right] = P\left[Z \geq \frac{K - \mu_1}{\sigma/\sqrt{n}}\right]$$

de donde resulta,

$$\frac{K - \mu_1}{\sigma/\sqrt{n}} = z_{1-\beta}, \qquad K = \mu_1 + z_{1-\beta}\frac{\sigma}{\sqrt{n}} \tag{2}$$

Resolviendo para n las ecuaciones (1) y (2), resulta

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2}{(\mu_0 - \mu_1)^2}.$$

NOTA.

Si la prueba unilateral es de cola a la derecha, esto es si se prueba:

$$H_0 : \mu = \mu_0 \text{ contra } H_1 : \mu = \mu_1, \text{ donde } \mu_0 < \mu_1$$

el tamaño n de la muestra requerida es también:

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2}{(\mu_0 - \mu_1)^2}.$$

Si la prueba es bilateral, esto es, si se prueba: $H_0 : \mu = \mu_0$ contra $H_1 : \mu = \mu_1$, $\mu_0 \neq \mu_1$, el tamaño n de la muestra requerida es

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{(\mu_0 - \mu_1)^2}.$$

EJEMPLO 10.4.

El gerente de producción de la empresa "PALMERA" que envasa conservas de palmito afirma que el peso neto del producto envasado tiene una media de 250 gramos y una desviación estándar de 24 gramos. Diseñe una prueba de hipótesis para la media de la población si se quiere que el riesgo sea 6.68% de rechazar tal afirmación cuando realmente es verdadera y el riesgo sea de 2.28% de aceptar la afirmación cuando la media de los pesos sea realmente 264 gramos.

SOLUCION.

Sean las hipótesis: H_0: μ=250 y H_1: μ=264 y sea K el valor crítico de la prueba.
Si se supone verdadera la hipótesis nula $H_0 : \mu = 250$, entonces la distribución de $Z = \dfrac{\overline{X} - 250}{24/\sqrt{n}}$ es normal $N(0,1)$, sólo si el tamaño de la muestra es grande.

Luego, para $\alpha = 0.0668 = P[\text{Error tipo I}]$, en esta distribución se tiene:

$$0.0668 = P[\text{rechazar } H_0 \ / \ H_0 : \mu = 250 \text{ es verdadera}]$$

$$0.0668 = P\left[\overline{X} > K/\mu = 250\right] = P\left[Z > \frac{K - 250}{24/\sqrt{n}}\right]$$

de donde resulta: $\dfrac{K - 250}{24/\sqrt{n}} = +1.5,$ $\qquad K = 150 + \dfrac{1.5 \times 24}{\sqrt{n}}$

También, si se supone verdadera la hipótesis alternativa $H_1 : \mu = 264$, entonces, la distribución de $Z = \dfrac{\overline{X} - 264}{24/\sqrt{n}}$ es normal $N(0,1)$.

Luego, para $\beta = 0.0228 = P[\text{Error tipo II}]$, en esta distribución se tiene:

$$0.0228 = P[\text{aceptar } H_0 \ / \ H_1 : \mu = 264\,]$$

$$0.0228 = P\left[\overline{X} \le K/\mu = 264\right] = P\left[Z \le \frac{K - 264}{24/\sqrt{n}}\right]$$

$$\frac{K - 264}{24/\sqrt{n}} = -2, \qquad K = 264 - \frac{2 \times 24}{\sqrt{n}}.$$

Luego de: $K = 250 + \dfrac{1.5 \times 24}{\sqrt{n}}$ y de $K = 264 - \dfrac{2 \times 24}{\sqrt{n}}$, se obtiene:

$\sqrt{n} = 6$, $\quad n$ =36, $\quad$ El valor crítico de la prueba es : $K = 256$.

Si $\overline{x}$ es un valor de media de la muestra de n =36 casos, se rechazará la hipótesis nula H_0 si $\overline{x} > 256$. En caso contrario, no se debe rechazar H_0.

10.3 Pruebas de hipótesis acerca de la media μ: Varianza σ^2 supuesta desconocida

A) Población no normal

Si la población no tiene distribución normal y si la varianza es desconocida, para probar hipótesis acerca de la media μ se utiliza la estadística Z (sólo si, el tamaño de la muestra es grande: $n \geq 30$):

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

cuya distribución es aproximadamente $N(0,1)$.

La desviación estándar σ se estima puntualmente por $\hat{s}$.

Luego, las regiones críticas de la pruebas de $H_0 : \mu = \mu_0$ contra cualquiera de las tres alternativas $H_1 : \mu > \mu_0$ ó $H_1 : \mu < \mu_0$ ó $H_1 : \mu \neq \mu_0$ son las mismas (aproximadamente) de la sección 10.2. (Ver **ejemplo 10.2** de esa sección)

B) Población normal

Si la población tiene distribución normal $N(\mu, \sigma^2)$, donde μ y σ^2 son parámetros desconocidas, para $n \geq 2$ la estadística de la prueba acerca de la media μ es:

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

cuya distribución es ***t*-Student** con $n-1$ grados de libertad.

Si se supone verdadera la hipótesis nula: $H_0 : \mu = \mu_0$, la estadística especificada por esta hipótesis es entonces, ahora:

$$T = \frac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}}$$

NOTA. La estructura de la prueba es **similar** que en el caso de σ conocida, salvo que el valor de σ se estima por $\hat{s}$ y la distribución normal estándar se sustituye por la distribución t de Student con $n-1$ grados de libertad.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, dado el nivel de significación α, en la distribución de $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n}) \sim t(n-1)$, se determinan los valores críticos $\pm t_{1-\alpha/2, n-1}$, tales que la probabilidad de rechazar H_0 cuando se supone verdadera sea (figura 10.6)

$$P[T < -t_{1-\alpha/2, n-1}] = \alpha/2 \quad \text{o} \quad P[T > t_{1-\alpha/2, n-1}] = \alpha/2 .$$

Luego, la ***región crítica* o de rechazo de H_0 *en el rango de variación de*** T es:

$$R.C. = \{T < -t_{1-\alpha/2, n-1} \;\; \text{o} \;\; T > t_{1-\alpha/2, n-1}\}$$

Regla de decisión: Se rechazará H_0 con riesgo α, si $t_k \in R.C.$ (o, si $t_k \notin R.A = \{-t_{1-\alpha/2, n-1} \leq T \leq t_{1-\alpha/2, n-1}\}$). No se rechazará H_0 en caso contrario.

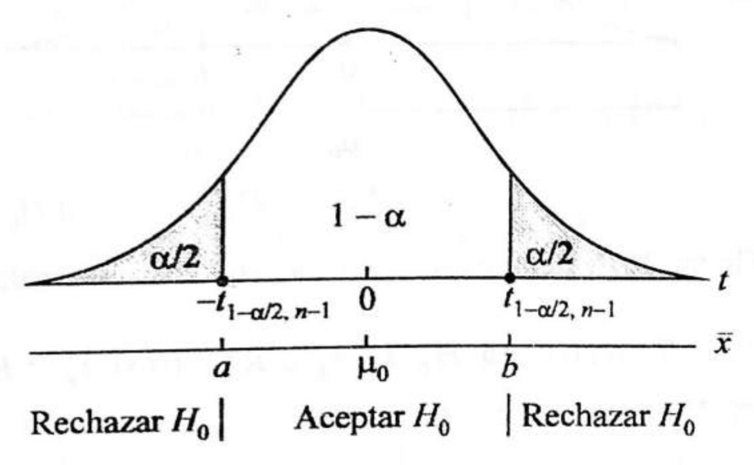


Figura 10.6: Región crítica bilateral en escalas t y $\overline{x}$

NOTA. Si se sustituye $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n})$ en $R.C$ se obtiene:
la ***región crítica* o de rechazo de H_0 *en el rango de variación de*** $\overline{X}$:

$$R.C. = \{\overline{X} < a \quad \text{o} \quad \overline{X} > b\}$$

donde $a = \mu_0 - t_{1-\alpha/2, n-1}(\hat{s}/\sqrt{n})$, y $b = \mu_0 + t_{1-\alpha/2, n-1}(\hat{s}/\sqrt{n})$

Regla de decisión: Siendo $\overline{x}$ el valor de $\overline{X}$ obtenido a partir de una muestra aleatoria de tamaño *n*, se rechazará H_0 con un riesgo α, si $\overline{x} \in R.C.$ (o si $\overline{x} \notin R.A. = (R.C.)^c$). No se rechazará H_0 en caso contrario (figura 10.6).

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu > \mu_0$, dado el nivel de significación α, en la distribución de $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n}) \sim t(n-1)$, se determina el valor $t_{1-\alpha, n-1}$ tal que: (figura 10.7)

$$P[T > t_{1-\alpha, n-1} / H_0 : \mu = \mu_0 \text{ verdadera}] = \alpha$$

Luego, ***la región crítica o de rechazo H_0 en el rango de variación de T*** es:

$$RC = \{T > t_{1-\alpha, n-1}\}$$

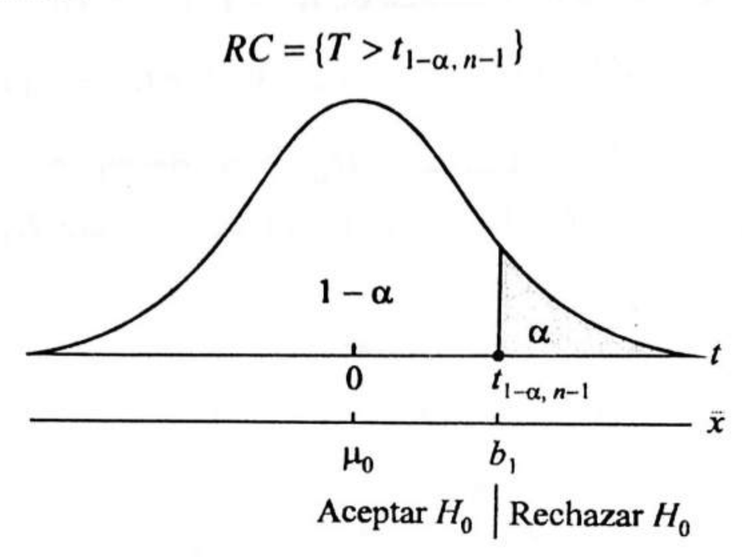


Figura 10.7: Región crítica cola a la derecha en escalas t y $\overline{X}$

Regla de decisión: Se rechazará H_0 si $t_k \in R.C.$ (o si $t_k \notin R.A. = \{T \leq t_{1-\alpha, n-1}\}$). No se rechazará H_0 en caso contrario.

NOTA. La región crítica ***o de rechazo H₀*** en $\overline{X}$ (figura 10.7) es: $RC = \{\overline{X} > b_1\}$, donde

$$b_1 = \overline{X} < \mu_0 + t_{1-\alpha, n-1}(\hat{s}/\sqrt{n})$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu < \mu_0$, dado el nivel de significación α, en la distribución de $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n}) \sim t(n-1)$ se determina el valor $t_{1-\alpha, n-1}$, tal que; (figura 10.8)

$$P[T < -t_{1-\alpha, n-1} / H_0 : \mu = \mu_0 \text{ verdadera}] = \alpha$$

Luego, *la región crítica **o de rechazo H_0** en el rango de variación de T* es:

$$RC = \{T < -t_{1-\alpha,\, n-1}\}$$

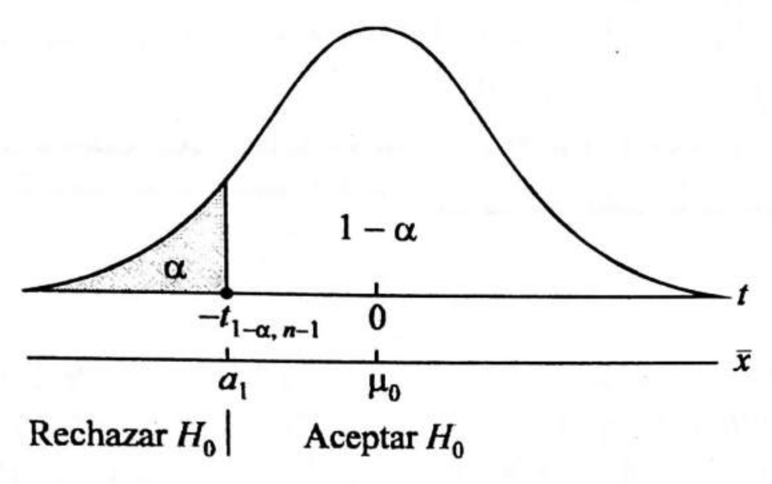


Figura 10.8: Región crítica cola a la izquierda en escalas t y $\overline{X}$

Regla de decisión: Se rechazará H_0 si $t_k \in R.C.$ (o si $t_k \notin R.A. = \{T \geq -t_{1-\alpha,\, n-1}\}$). No se rechazará H_0 en caso contrario (figura 10.8).

NOTA. La región crítica ***o de rechazo H_0*** en $\overline{X}$ (figura 10.8) es $RC = \{\overline{X} < a_1\}$, donde:

$$a_1 = \overline{X} < \mu_0 - t_{1-\alpha,\, n-1}(\hat{s}/\sqrt{n})$$

EJEMPLO 10.5.

Los registros de un propietario de una estación de combustible indican que la media del número de galones de gasolina que vende a sus clientes es igual a 4 galones. Además, los registros muestran que los consumos de gasolina de sus clientes tienen una distribución normal. Sin embargo, debido a la reciente alza en el precio de la gasolina se cree que este consumo ha bajado. Para verificar esta hipótesis se escogió una muestra aleatoria de 15 de sus clientes resultando los siguientes consumos de gasolina en galones:

4.25, 3.75, 4.05, 3.8, 35, 4, 3.75,
2.5, 6.1, 2.5, 2.5, 3.4, 3.2, 2.8, 5

con estos datos y con un nivel de significación de 0.05, ¿el incremento en el precio de la gasolina ha influido en la baja del consumo promedio?. ¿Cuál es el valor de la probabilidad P?.

SOLUCION.

Sea X la variable que representa el consumo en galones de gasolina, cuya distribución es normal con media μ y varianza σ^2 desconocidas.

1. *Hipótesis*: $H_0 : \mu = 4$ contra $H_1 : \mu < 4$
2. *Nivel de significación* $\alpha = 0.05$
3. *Estadística*: Población normal, con varianza desconocida y $n = 15$. Si $H_0 : \mu = 4$ es verdadera, la estadística es

$$T = \frac{\overline{X} - 4}{\hat{s}/\sqrt{n}}$$

que se distribuye según una *t*-Student con 14 grados de libertad.

4. *Región crítica*: Con el nivel de significación $\alpha = 0.05$ y para una prueba de hipótesis unilateral cola a la izquierda, en la tabla de probabilidades de *t*-Student se halla el valor crítico: $t_{0.95,14} = 1.761$.

   Consecuentemente, la región crítica o de rechazo de H_0 es: $RC = \{T < -1.761\}$
5. *Cálculos*: De los datos de la muestra se obtiene:

$$n = 15,\ \bar{x} = 3.6733,\ \hat{s} = 0.9812,\ \text{error estándar: } \hat{s}/\sqrt{n} = 0.2534$$

$$t_k = \frac{\bar{x} - 4}{\hat{s}/\sqrt{n}} = \frac{3.6733 - 4}{0.2534} = -1.289.$$

6. *Decisión*: Dado que $t_k = -1.289 \notin R.C$, deberíamos aceptar H_0 y concluir que el incremento en el precio de la gasolina no bajó el consumo promedio.

Utilizando una calculadora o el paquete de computo estadístico *MCEST*, se encuentra el valor de la probabilidad: $P = P[T < -1.289] = 0.109$.

NOTA. El valor crítico de la prueba en $\overline{X}$ es $K = 4 - 1.761 \times 0.2534 = 3.554$, por tanto, dado que $\bar{x} = 3.6733 \notin R.C = \{\overline{X} < 3.554\}$, deberíamos aceptar H_0.

EJEMPLO 10.6.

La compañía PROANZA produce un cable de acero que tiene una resistencia media a la ruptura de 300 kg.. La compañía estudia la posibilidad de usar un proceso nuevo y más barato. Lo hará si estadísticamente se comprueba que el cable producido con el nuevo proceso tiene una resistencia media a la ruptura igual o mayor que 300 kg.. Para esto se escogió una muestra aleatoria de 36 cables producidos con el nuevo proceso resultando una media de resistencia a la ruptura de 296.25 kg. y una desviación estándar $\hat{s} = 15$ kg.. ¿Debería esta compañía adoptar el nuevo proceso, si está dispuesto a asumir un error tipo I de 0.05?.

Resuelva el problema si la distribución de la resistencia a la ruptura es:

a) Normal

b) Desconocida no normal.

SOLUCION.

Sea X la resistencia a la ruptura del cable
Se debe probar la hipótesis nula $H_0 : \mu \geq 300$ contra $H_1 : \mu < 300$ al nivel de significación del 5%, y usando una muestra de tamaño: $n = 36$.

a) Si X tiene distribución normal con varianza σ^2 desconocida y si $H_0 : \mu = 300$ es supuesta verdadera, la estadística de la prueba es t-student:

$$T = \frac{\overline{X} - 300}{\hat{s}/\sqrt{n}} \sim t(35)$$

Para $\alpha = 0.05$ y una prueba unilateral de cola a la izquierda, en la distribución $t(35)$ se encuentra el valor crítico: $t_{0.950,35} = -1.69$.

Luego, la región crítica es: $RC = \{T < -1.69\}$

De los datos de la muestra, se tiene:

$$n = 36,\ \bar{x} = 296.25.,\ \hat{s} = 15,\ \text{error estándar: } \hat{s}/\sqrt{n} = 2.5$$

$$t_k = \frac{\bar{x} - 300}{\hat{s}/\sqrt{n}} = \frac{296.25 - 300}{2.5} = -1.5.$$

Dado que, $t_k = -1.5 \notin R.C.$, se debería rechazar H_0.

Observe que la significación de la prueba es: $P = P[t_{35} <- 1.5] = 0.0713$

b) Si X no se distribuye en forma normal con varianza σ^2 desconocida, con $n \geq 30$ se hace $\sigma = \hat{s}$, y si $H_0 : \mu = 300$ es supuesta verdadera, la estadística es:

$$Z = \frac{\overline{X} - 300}{\hat{s}/\sqrt{n}}$$

que se distribuye aproximadamente normal $N(0,1)$.

Para el nivel de significación $\alpha = 0.05$ y para una prueba unilateral de cola a la derecha, en la distribución de Z se encuentra: $z_{1-\alpha} = z_{0.95} = 1.645$.

Luego, la región crítica de la prueba es: $RC = \{Z < -1{,}645\}$.

De la muestra aleatoria resulta:

$$z_k = \frac{\bar{x} - 300}{\hat{s}/\sqrt{n}} = \frac{296.25 - 300}{2.5} = -1.5.$$

Ya que $z_k = -1.5 \notin R.C.$, debemos aceptar H_0 y concluir que conviene a la compañía adoptar el nuevo proceso de producir el cable.

Observe que la significación de la prueba es: $P = P[Z < -1.5] = 0.0668$.

10.4 Pruebas de hipótesis acerca de una varianza

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño *n*, seleccionada de una población normal con varianza σ^2, parámetro desconocido Y sea la varianza muestral:

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

Entonces, la variable aleatoria:

$$X = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

tiene distribución chi-cuadrado con $n-1$ grados de libertad. Esta estadística se utiliza para probar hipótesis acerca de una varianza.

Si se supone verdadera la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$, la estadística es:

$$X = \frac{(n-1)\hat{S}^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Su valor $x_k = \dfrac{(n-1)\hat{s}^2}{\sigma_0^2}$ que resulta de la muestra aleatoria, se utiliza para la prueba de H_0, contra una alternativa unilateral o bilateral.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \sigma^2 = \sigma_0^2$ contra $H_1 : \sigma^2 \neq \sigma_0^2$, dado un nivel de significación α, en la distribución $\chi^2(n-1)$ se hallan los valores críticos $\chi^2_{\alpha/2,n-1}$ y $\chi^2_{1-\alpha/2,n-1}$` (figura 10.9), tales que la probabilidad de rechazar la hipótesis nula H_0 cuando es supuesta verdadera, es igual a:

$$P[X < \chi^2_{\alpha/2,n-1}] = \alpha/2 \quad \text{o} \quad P[X > \chi^2_{1-\alpha/2,n-1}] = \alpha/2.$$

La *Región crítica* o de rechazo de H_0, es entonces, el intervalo

$$R.C. = \{X < \chi^2_{\alpha/2,n-1} \quad \text{o} \quad X > \chi^2_{1-\alpha/2,n-1}\}.$$

La *Regla de decisión* es: Rechazar H_0 con un riesgo α, si $x_k \in R.C.$ (o si $x_k \notin R.A. = [\chi^2_{\alpha/2,\,n-1}, \chi^2_{1-\alpha/2,\,n-1}]$). No rechazar H_0 en caso contrario.

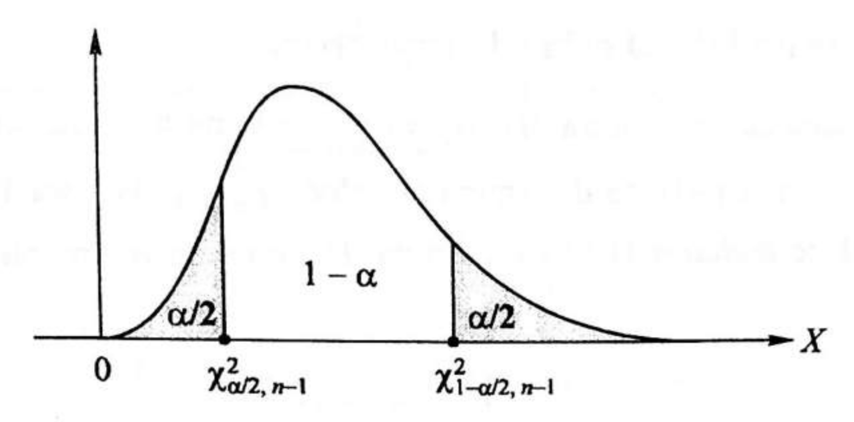


Figura 10.9: Región crítica para la prueba de $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 \neq \sigma_0^2$

2) Contraste unilateral de cola a la derecha..

Si se prueba $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 > \sigma_0^2$, dado un nivel de significación α, en la distribución $\chi^2(n-1)$ se determina el valor $\chi^2_{1-\alpha,\,n-1}$ (figura 10.10) tal que la probabilidad de rechazar la hipótesis nula H_0 cuando realmente es verdadera es igual a:

$$P[X > \chi^2_{1-\alpha,\,n-1}] = \alpha.$$

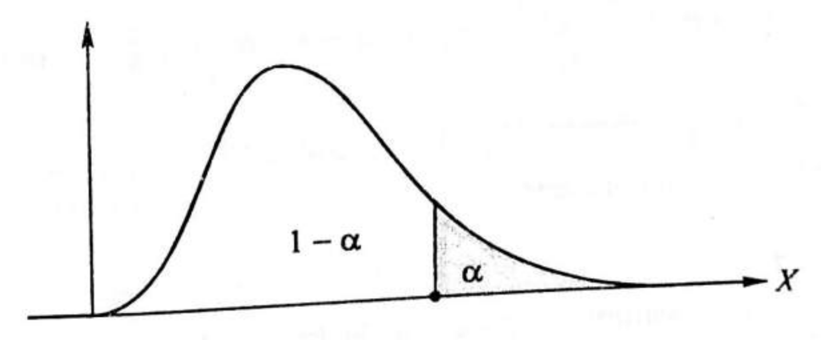


Figura 10.10: Región crítica para la prueba de $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 > \sigma_0^2$

Luego, la *región crítica* o de rechazo de H_0 es el intervalo:

$$R.C. = \{X > \chi^2_{1-\alpha,\,n-1}\}.$$

La *regla de decisión* es: rechazar H_0 al nivel α si $x_k \in R.C.$ (o si $x_k \notin R.A. = \{X \leq \chi^2_{1-\alpha,n-1}\}$). No rechazar H_0 en caso contrario.

3) Contraste unilateral cola a la izquierda.

Si la prueba es de contra $H_1 : \sigma^2 < \sigma_0^2$, dado un nivel de significación α, en la distribución $\chi^2(n-1)$ se determina el valor $\chi^2_{\alpha,n-1}$ (figura 10.11) tal que la probabilidad de rechazar la hipótesis nula H_0 cuando realmente es verdadera es igual a:

$$P[X < \chi^2_{\alpha,n-1}] = \alpha.$$

Luego, la *región crítica* o de rechazo de H_0 es el intervalo:

$$R.C. = \{X < \chi^2_{\alpha,n-1}\}$$

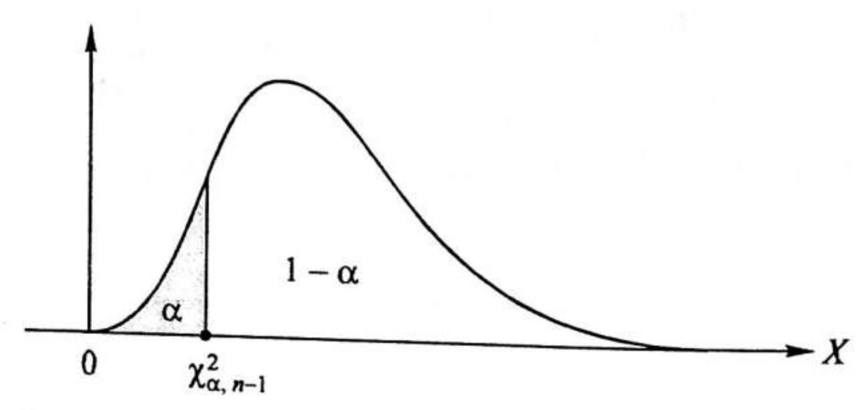


Figura 10.11: Región crítica para la prueba de H_0: $\sigma^2 = \sigma_0^2$ contra H_1: $\sigma^2 < \sigma_0^2$

Regla de decisión: rechazar H_0 si $x_k \in R.C.$ (o si $x_k \notin R.A. = \{X \geq \chi^2_{\alpha,n-1}\}$). No rechazar H_0 en caso contrario.

EJEMPLO 10.7.

Un fabricante de tornillos asegura que la longitud de un tipo especial de sus tornillos de alta precisión tiene distribución normal con una desviación estándar igual a 0.2 milímetros. Una muestra aleatoria de diez de estos tornillos ha dado las siguientes longitudes en milímetros.

20.50	20.80	20.26	20.72	20.69
20.37	20.70	20.32	20.96	20.40

Al nivel de significación de 0.05, ¿se justifica la afirmación que la desviación estándar verdadera es 0.2 milímetros?

SOLUCION.

1. *Hipótesis*: $H_0 : \sigma^2 = (0.2)^2$ contra $H_1 : \sigma^2 \neq (0.2)^2$

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Población normal, con $n = 10$, y suponiendo verdadera la hipótesis nula $H_0 : \sigma^2 = (0.2)^2$, la estadística es

$$X = \frac{(n-1)\hat{S}^2}{(0.2)^2}$$

que se distribuye como chi-cuadrado con 9 grados de libertad.

4. *Región crítica*: Para $\alpha = 0.05$ y para un contraste bilateral, en la tabla chi-cuadrado se encuentran los siguientes valores críticos:

Lado izquierdo: $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,9} = 2.70$

Lado derecho: $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,9} = 19.02$.

Luego, la región crítica es: $R.C. = \{X < 2.70 \text{ o } X > 19.02\}$.

5. *Cálculos*: De los datos de la muestra resulta:

$$\hat{s}^2 = 0.05457,$$

entonces,

$$x_k = \frac{(n-1)\hat{s}^2}{(0.2)^2} = \frac{9\hat{s}^2}{(0.2)^2} = \frac{9 \times 0.05457}{(0.2)^2} = 12.278.$$

6. *Decisión*: Como $x_k = 12.278 \notin R.C.$ no se debería rechazar H_0 y concluimos que la desviación estándar es igual a 0.2 mm.

NOTA. La probabilidad $P = P[\chi^2(9) > 12.278] = 0.1981$

10.5 Pruebas de hipótesis acerca de la razón de dos varianzas

Sean $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos muestras aleatorias *independientes* de tamaños respectivos n_1 y n_2, escogidas de **dos poblaciones normales** con **varianzas** respectivas σ_1^2 y σ_2^2. Entonces, la estadística,

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución de probabilidad F con grados de libertad $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$. Esto es, $F \sim F(r_1, r_2)$. Esta estadística se utiliza para probar igualdad de varianzas.

Si se supone verdadera la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ o $\sigma_1^2 / \sigma_2^2 = 1$ la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} \sim F(r_1, r_2)$$

Su valor $f_k = \dfrac{\hat{s}_1^2}{\hat{s}_2^2}$ que resulta de dos muestras aleatorias, se utiliza para probar la hipótesis nula H_0 contra cualquier alternativa unilateral o bilateral

Observe que para obtener la estadística F, no se requiere asumir que las dos poblaciones tengan igual promedio.

1) Prueba bilateral.

Si se prueba $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$, dado un nivel de significación α, en la distribución de $F(r_1, r_2)$ se encuentran los valores $f_{\alpha/2, r_1, r_2}$ y $f_{1-\alpha/2, r_1, r_2}$ tales que la probabilidad de rechazar H_0 cuando realmente es verdadera es igual a (ver figura 10.12).

$$P[F < f_{\alpha/2, r_1, r_2}] = \alpha/2 \quad \text{o} \quad P[F > f_{1-\alpha/2, r_1, r_2}] = \alpha/2.$$

Luego la *región crítica* o de rechazo de H_0 es

$$R.C. = \{F < f_{\alpha/2, r_1, r_2} \text{ o } F > f_{1-\alpha/2, r_1, r_2}\}.$$

La *regla de decisión* es rechazar H_0 si $f_k \in R.C.$ o si $f_k \notin R.A. = [f_{\alpha/2, r_1, r_2}, f_{1-\alpha/2, r_1, r_2}]$. No rechazar H_0 en caso contrario.

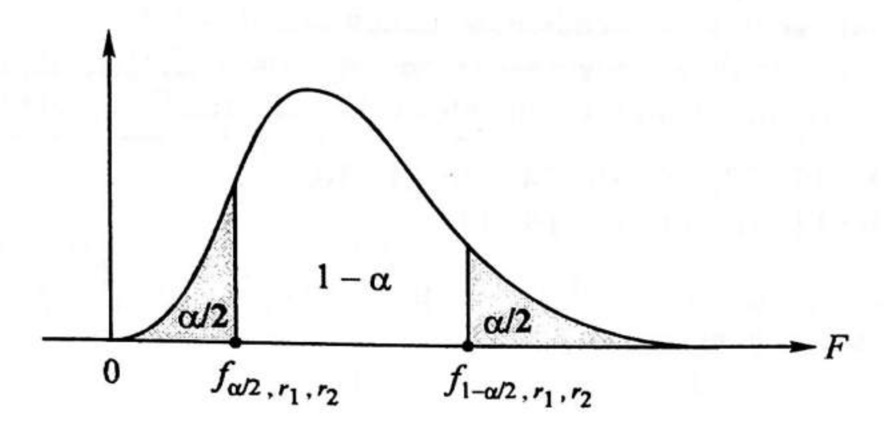


Figura 10.12: Región crítica de la prueba de H_0: $\sigma_1^2 = \sigma_2^2$ contra H_1: $\sigma_1^2 \neq \sigma_2^2$

2) Prueba unilateral cola derecha.

Si se prueba $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 > \sigma_2^2$, dado el nivel de significación α, en la distribución $F(r_1, r_2)$ se determina el valor $f_{1-\alpha, r_1, r_2}$ tal que la probabilidad de rechazar H_0 cuando realmente es verdadera es igual a;

$$P[F > f_{1-\alpha, r_1, r_2}] = \alpha.$$

Luego, la región crítica o de rechazo de H_0 es: $R.C. = \{F > f_{1-\alpha, r_1, r_2}\}$

La *regla de decisión* es: Rechazar H_0, si $f_k \in R.C.$ o si $f_k \notin R.A. =]-\infty, f_{1-\alpha, r_1, r_2}]$. No rechazar H_0 en caso contrario

3) Prueba unilateral cola izquierda.

Si se prueba $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 < \sigma_2^2$, dado el nivel de significación α, en la distribución $F(r_1, r_2)$ se determina el valor f_{α, r_1, r_2} tal que la probabilidad de rechazar H_0 cuando realmente es verdadera es igual a;

$$P[F < f_{\alpha, r_1, r_2}] = \alpha$$

Luego, la región crítica o de rechazo de H_0 es: $R.C. = \{F < f_{\alpha, r_1, r_2}\}$.

La *regla de decisión* es: Rechazar H_0, si $f_k \in R.C.$ o si $f_k \notin R.A. = \{F \geq f_{\alpha, r_1, r_2}\}$. No rechazar H_0 en caso contrario.

EJEMPLO 10.8.

Un corredor de valores de la bolsa de Lima estudia las porcentajes de rendimiento de las empresas del sector minero y del sector financiero. Se sabe que tasas de los rendimientos independientes tienen distribución normal. Dos muestras aleatorias de las tasas de 8 empresas del sector minero (M) y de 6 empresas del sector financiero (F) han dado los siguientes valores de rendimiento en porcentajes:

Sector M: 17, 23, 25, 18, 24, 20, 21, 16.

Sector F : 13, 16, 14, 12, 15, 14.

Con un nivel de significación de 0.05, ¿se puede concluir que hay mas variación en los valores del sector minero?.

SOLUCION.

Sean X_1 y X_2 las variables que representan los valores de rendimiento de los sectores M y F respectivamente. Estas variables tienen distribuciones normales con varianzas desconocidas respectivas: σ_1^2 y σ_2^2.

De la muestra se obtiene: $\hat{s}_1^2 = 11.1429$ y $\hat{s}_2^2 = 2$. Como la varianza de la muestra M es mayor que la varianza de la muestra F, es razonable plantear una prueba unilateral.

1. *Hipótesis*: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 > \sigma_2^2$.

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Poblaciones normales y suponiendo verdadera la hipótesis nula H_0, para $n_1 = 8$ y $n_2 = 6$, la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

que se distribuye como $F(7,5)$.

4. *Región crítica*. Para $\alpha = 0.05$ y la prueba unilateral cola a la derecha, en la distribución de $F(7,5)$, la región crítica es

$$R.C. = \{ F > 4.88 \}.$$

5. *Cálculos*. De los datos de la muestra se obtiene:

$$\hat{s}_1^2 = 11.1429, \quad \hat{s}_2^2 = 2 \quad \text{y} \quad f_k = \hat{s}_1^2 / \hat{s}_2^2 = 5.5714.$$

6. *Decisión*. Como $f_k = 5.5714 \in R.C.$ se debería rechazar H_0 al nivel 0.05 y concluir que los valores del sector F tienen menor variación .

NOTA La significación de la prueba es: $P[F > 5.5714] = 0.0383$

10.6 Pruebas de hipótesis acerca de dos medias

10.6.1 Pruebas de hipótesis acerca de dos medias: Varianzas σ_1^2 y σ_2^2 supuestas conocidas

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones independientes, con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 respectivas supuestas conocidas.

Si las *dos poblaciones son normales*, entonces, las estadísticas $\overline{X}_1$ y $\overline{X}_2$ tienen respectivamente distribución normal $N(\mu_1, \sigma_1^2/n_1)$ y $N(\mu_2, \sigma_2^2/n_2)$ para $n_1 \geq 2$ y $n_2 \geq 2$. Luego, la estadística $\overline{X}_1 - \overline{X}_2$ tiene distribución exactamente normal: $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Si las dos *poblaciones no* son *normales* pero n_1 y n_2 son suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$), entonces, $\overline{X}_1 - \overline{X}_2$ tiene distribución aproximadamente normal: $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Luego, según sean las dos poblaciones normales o no, la estadística

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

tiene distribución exactamente o aproximadamente normal $N(0,1)$.

Si suponemos *verdadera* la hipótesis nula $H_0 : \mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$, la estadística es:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

Su valor $z_k = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ que resulta de dos muestras, se utiliza para probar H_0 contra cualquiera de las hipótesis alternativas $H_1 : \mu_1 \neq \mu_2$ ó $H_1 : \mu_1 > \mu_2$ ó $H_1 : \mu_1 < \mu_2$.

La estructura de la prueba es similar a los casos descritos en la sección 10.2, usando la distribución de Z.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, la región crítica o de rechazo de H_0 **en el rango de variación** de Z es:

$$R.C. = \{Z < -z_{1-\alpha/2} \text{ o } Z >_{1-\alpha/2}\}.$$

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$, la región crítica o de rechazo de H_0 en la variación de Z es,

$$R.C. = \{Z > z_{1-\alpha}\}.$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 < \mu_2$, la región crítica o de rechazo de H_0 en los valores de Z es:

$$R.C. = \{Z < -z_{1-\alpha}\}.$$

NOTA. Cuando las hipótesis son de la forma

1) $H_0 : \mu_1 - \mu_2 = d_0$ contra $H_1 : \mu_1 - \mu_2 \neq d_0$
2) $H_0 : \mu_1 - \mu_2 = d_0$ contra $H_1 : \mu_1 - \mu_2 > d_0$
3) $H_0 : \mu_1 - \mu_2 = d_0$ contra $H_1 : \mu_1 - \mu_2 < d_0$

La estadística de la prueba es,

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d_0}{\sqrt{\left(\sigma_1^2/n_1\right) + \left(\sigma_2^2/n_2\right)}}.$$

cuya distribución es exactamente o aproximadamente normal $N(0,1)$, según sean las dos poblaciones normales o no.

NOTA. Si las dos *poblaciones no* son *normales* y las varianzas σ_1^2 **y** σ_2^2 **son desconocidas** el uso de la estadística Z es válida haciendo $\sigma_1^2 = \hat{s}_1^2$ y $\sigma_2^2 = \hat{s}_2^2$, solo si n_1 y n_2 son suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$),

EJEMPLO 10.9.

El gerente de ventas de la gran compañía C&P analiza dos técnicas de ventas A y B. Escogió dos muestras aleatorias independientes de 50 vendedores. La primera, aplicó la técnica A y la segunda, la técnica B. Al final de un mes el número de ventas por vendedor ha dado las medias respectivas de 67 y 60 y las varianzas respectivas 225 y 100.

a) Al nivel de significación del 5%, ¿presentan los resultados muestrales suficiente evidencia que indique que la técnica A da mejores resultados que la técnica B?.

b) Al nivel de significación del 5%, ¿se puede inferir que la media de la técnica A es mayor que la media de B en más de 2?.

c) ¿En cuánto deberían incrementarse los tamaños de las muestras que se escogen para cada técnica de manera que una diferencia observada igual a 5 en las ventas medias muestrales de A menos B sea significativa al nivel α=1.5%?

SOLUCION.

Sean X_1 y X_2 las ventas obtenidas con las técnicas A y B respectivamente y μ_1 y μ_2 sus medias respectivas.

Se desconocen las distribuciones de probabilidades de X_1 y X_2, pero las muestras son grandes.

a) 1. *Hipótesis*: $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Si se supone verdadera la hipótesis H_0 y para muestras grandes, la estadística apropiada es:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\hat{s}_1^2/n_1\right) + \left(\hat{\sigma}_2^2/n_2\right)}}$$

cuya distribución es aproximadamente normal estándar $N(0,1)$.

4. *Región crítica*: Para $\alpha = 0.05$ y una prueba unilateral de cola a la derecha, en la distribución de Z se encuentra el valor $z_{0.9500} = 1.645$. Luego, la región crítica o de rechazo de H_0 es,

$$R.C. = \{Z > 1.645\}.$$

5. *Cálculos*: De los datos se tiene

$$n_1 = n_2 = 50, \quad \bar{x}_1 = 67, \quad \bar{x}_2 = 60, \quad \hat{s}_1 = 15, \quad \hat{s}_2 = 10$$

$$E.S\text{=Error estándar} = \sqrt{\left(\hat{s}_1^2/n_1\right) + \left(\hat{\sigma}_2^2/n_2\right)} = 2.5495$$

$$z_k = \frac{\bar{x}_1 - \bar{x}_2}{ES} = \frac{67-60}{2.5495} = 2.7456$$

6. *Decisión*: Ya que $z_K = 2.7456 \in R.C.$ al nivel 5%, debemos rechazar H_0 y concluir que la técnica A da mayores resultados que la B.

Observe que la significación es: $P = P[Z > 2.7456] = 0.0030$.

b) Se debe probar $H_0: \mu_1 - \mu_2 = 2$ contra $H_1: \mu_1 - \mu_2 > 2$.
Si H_0 es verdadera, la estadística de la prueba es:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0,1)$$

La región de rechazo de H_0 de la prueba unilateral cola derecha al nivel 0.05 es la misma del caso a), esto es:

$$R.C. = \{Z > 1.645\}.$$

$$z_k = \frac{(\bar{x}_1 - \bar{x}_2) - 2}{ES} = \frac{(67-60)-2}{2.5495} = 1.961$$

Ya que $z_k = 1.961 \in R.C.$, debemos rechazar H_0 y concluir con riesgo tipo I de 0.05 que la media de la técnica A es mayor a la media de la B en más de 2.

c) Sea n el tamaño de cada una de las dos muestras tomadas de las técnicas A y B. Si la hipótesis nula H_0 es verdadera, y si además

$$\bar{x}_1 - \bar{x}_2 = 5,$$

se tiene

$$z_k = \frac{7}{\sqrt{\dfrac{15^2 + 10^2}{n}}} = 0.2773\sqrt{n},$$

Para $\alpha = 0.015$ y una prueba unilateral de cola a la derecha, en la distribución de Z se halla:

$$z_{1-\alpha} = z_{0.985} = 2.17$$

La región crítica es: $R.C = \{Z > 2.17\}$

La diferencia observada de 5 en las medias de las muestras será significativa al nivel de 1.5%, si

$$0.2773\sqrt{n} \in R.C.$$

Esto es, si $0.2773\sqrt{n} > 2.17$, $\sqrt{n} > 7.82546$, $n \geq 62$.

De aquí que se debe incrementar cada muestra en al menos: $62 - 50 = 12$ casos

10.6.2 Pruebas de hipótesis acerca de dos medias: Varianzas σ_1^2 y σ_2^2 supuestas desconocidas

A) Poblaciones no normales

Si las **dos muestras aleatorias independientes** de tamaños n_1 y n_2 se seleccionan respectivamente de dos poblaciones cuyas *distribuciones son no normales* con varianzas σ_1^2 y σ_2^2 supuestas desconocidas , entonces, siempre que los tamaños de las muestras sean grandes; $n_1 \geq 30$ y $n_2 \geq 30$, los parámetros σ_1 y σ_2 se estiman respectivamente por $\hat{s}_1$ y $\hat{s}_2$. En este caso, para probar la hipótesis nula $H_0 : \mu_1 - \mu_2 = 0$ contra una alternativa bilateral o unilateral, se utiliza la estadística Z dada por:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}}$$

que se distribuye aproximadamente normal $N(0,1)$.

Las regiones críticas y las reglas de decisión para las pruebas de la hipótesis nula $H_0 : \mu_1 - \mu_2 = 0$ (ó $H_0 : \mu_1 - \mu_2 = d_0$) contra cualquiera de las hipótesis alternativas $H_1 : \mu_1 \neq \mu_2$ ó $H_1 : \mu_1 > \mu_2$ ó $H_1 : \mu_1 < \mu_2$ **son las mismas que con varianzas conocidas de la sección 10.6.1**. Ver ejemplo 10.9

B) Poblaciones normales

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias y $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de **dos muestras aleatorias independientes** de tamaños n_1 y n_2 respectivamente seleccionadas de *dos poblaciones normales* con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 desconocidas.

B1) Varianzas desconocidas supuesta iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Si las poblaciones son normales, independientes, y con varianzas desconocidas supuestas iguales, entonces, la estadística:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_c^2}{n_1} + \frac{\hat{S}_c^2}{n_2}}}$$

tiene distribución *t*-student con $n_1 + n_2 - 2$ grados de libertad, en donde la varianza común:

$$\hat{S}_c^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

es un *estimador insesgado de la varianza común* σ^2.

Si la hipótesis nula $H_0 : \mu_1 = \mu_2$ es verdadera, entonces, la estadística *T* es:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\hat{S}_c^2}{n_1} + \frac{\hat{S}_c^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

Su valor $t_k = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

que resulta de dos muestra aleatorias, se utiliza para probar H_0 contra una de las hipótesis alternativas bilateral o unilateral .

Las regiones críticas y las reglas de decisión de esta prueba *t*, son similares a las de la sección 10.3 usando la distribución *t* con $n_1 + n_2 - 2$ grados de libertad.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{T < -t_{1-\alpha/2,\ n_1+n_2-2} \text{ o } T > t_{1-\alpha/2,\ n_1+n_2-2}\}$$

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{T > t_{1-\alpha,\ n_1+n_2-2}\}$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 < \mu_2$, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{T < -t_{1-\alpha,\ n_1+n_2-2}\}$$

B2) Varianzas desconocidas supuestas distintas $\sigma_1^2 \neq \sigma_2^2$

Si las varianzas de las dos poblaciones normales independientes son desconocidas supuestas diferentes, entonces, la estadística

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}},$$

tiene distribución t-student con r grados de libertad, siendo

$$r = \frac{\left[\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2\right]^2}{\frac{\left[\hat{s}_1^2/n_1\right]^2}{n_1 - 1} + \frac{\left[\hat{s}_2^2/n_2\right]^2}{n_2 - 1}}$$

si r no es entero, se redondea al entero más cercano.

Si la hipótesis nula $H_0 : \mu_1 = \mu_2$ se supone verdadera, entonces

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}} \sim t(r)$$

Su valor $$t_k = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}}$$

que resulta de **dos muestras aleatorias independientes**, se utiliza para probar H_0 contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión de esta prueba t, son similares a las del caso B1 usando la distribución t con r grados de libertad.

EJEMPLO 10.10.

Una firma comercializadora está interesada en vender arroz embolsado por kilos que tenga el menor porcentaje de granos quebrados. Recibe el informe de dos molineras A y B que afirman tener el mejor arroz embolsado con el más bajo porcentaje de granos quebrados por kilo. Para tomar la decisión estadística se seleccionó una muestra aleatoria de 11 y otra de 10 bolsas de arroz de un kilo de

las molineras A y B respectivamente resultando los siguientes porcentajes de granos quebrados por kilo:

A: 2, 4, 6, 7, 7, 7, 8, 8, 9, 9, 10.

B: 2, 2, 3, 3, 3, 4, 4, 5, 6, 8

Se sabe que las poblaciones independientes de granos quebrados por kilo se distribuyen según el modelo de probabilidad normal.

a) Con un nivel de significación de 0.05, ¿se puede concluir que las varianzas poblacionales son iguales?
b) Con un nivel de significación de 0.05, ¿ se puede concluir que son iguales las medias de los porcentajes de granos quebrados por kilos de las molineras A y B?
c) ¿Qué tipo de arroz se recomienda comercializar a la firma?.

SOLUCION.

Sean X_1 y X_2 las variables aleatorias que representan los porcentajes de granos quebrados por kilo de las molineras A y B respectivamente. Se sabe que $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$ con parámetros desconocidos.

a) Prueba de la homogeneidad de varianzas

1. *Hipótesis*: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Poblaciones normales. Suponiendo verdadera la hipótesis nula H_0, para $n_1 = 11$ y $n_2 = 10$, la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

que se distribuye como $F(10, 9)$.

4. *Región crítica*. Para el nivel de significación $\alpha = 0.05$ y una prueba bilateral, en la distribución de $F(10, 9)$ se encuentran:

$$f_{0.975,10,9} = 3.96 \quad \text{y} \quad f_{0.025,10,9} = 1/f_{0.975,9,10} = 1/3.78 = 0.265$$

Luego, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{F < 0.265 \quad \text{o} \quad F > 3.96\}$$

5. *Cálculos*. De los datos de la muestra se obtiene:

$$\hat{s}_1^2 = 5.4, \quad \hat{s}_2^2 = 3.556 \quad \text{y} \quad f_k = \hat{s}_1^2 / \hat{s}_2^2 = 1.519.$$

6. *Decisión*. Como $f_k = 1.519 \notin R.C.$ Se debería aceptar H_0 y concluir que las varianzas de A y B son iguales.

b) Prueba de la diferencia de las dos medias.

1. *Hipótesis*: $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística de la prueba*: Si se supone H_0 verdadera y dado que hay prueba de que las varianzas poblacionales son iguales, la estadística apropiada es:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\hat{S}_c^2}{n_1} + \dfrac{\hat{S}_c^2}{n_2}}}$$

que se distribuye según una t–Student con grados de libertad: $n_1 + n_2 - 2 = 19$.

4. *Región crítica*: Para $\alpha = 0.05$ y una prueba de hipótesis bilateral, en la distribución $t(19)$ se encuentra $t_{0.975,19} = 2.093$. La región de rechazo de H_0 en la variación de T es el intervalo:

$$R.C. = \{\, T < -2.093 \;\; \text{o} \;\; T > 2.093 \}.$$

5. *Cálculos*: De los datos se tiene:

$$n_1 = 11,\ \overline{x}_1 = 7,\ \hat{s}_1 = 2.3238,$$
$$n_2 = 10,\ \overline{x}_2 = 4,\ \hat{s}_2 = 1.8856$$

$$\hat{s}_c^2 = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2} = \frac{10 \times 5.4 + 9 \times 3.556}{11 + 10 - 2} = 4.5265$$

$$ES = \text{Error estándar de la diferencia} = \sqrt{\frac{\hat{s}_c^2}{n_1} + \frac{\hat{s}_c^2}{n_2}} = 0.9296$$

$$t_k = \frac{\overline{x}_1 - \overline{x}_2}{ES} = \frac{7 - 4}{0.9296} = 3.227$$

6. *Decisión*: Ya que $t_k = 3.227 \in R.C.$, debemos rechazar H_0 y concluir que las medias de los porcentajes de granos quebrados de las dos molineras son diferentes.

c) Como las medias de las dos poblaciones son diferentes, planteamos las hipótesis:

$H_0 : \mu_1 = \mu_2$ (Ambos tipos de arroz son iguales).

$H_1 : \mu_1 > \mu_2$ (Arroz tipo B es mejor que el tipo A).

Con $\alpha = 0.05$ y 19 grados de libertad, para la prueba unilateral de cola a la izquierda se encuentra el valor crítico: $t_{0.95,19} = 1.729$. Luego, la región crítica es: $R.C. = \{T > 1.729\}$.

Como $t_k = 3.227 \in R.C.$, debemos rechazar H_0 y concluir que el mejor arroz es de tipo B, ya que el promedio de los porcentajes de granos quebrados por kilo es menor que el de A.

NOTA. El lector debería resolver este problema utilizando una calculadora con estadística avanzada o un paquete de computo. Por ejemplo, con el paquete *MCEST* **para la prueba de varianzas** en la distribución $F(10, 9)$ se obtiene: $P[F>1.519]=0.2707$. Luego, 0.2707 es la significación para una prueba unilateral y $2\times0.2707=0.5414$ para una prueba bilateral. Por lo que se infiere que las varianzas poblacionales son iguales a cualquier nivel de significación dado.

También, **para la prueba de dos medias** en la distribución $t(19)$, se obtiene: $P[T>3.227]=0.0022$. Luego, 0.0022 **es la significación** para una prueba unilateral y $2\times0.0022=0.0044$ para una prueba bilateral. Por lo que se debe rechazar H_0 en una prueba unilateral o bilateral a cualquier nivel de significación dado.

EJEMPLO 10.11.

Un analista compara dos métodos de enseñanza de Matemática básica: El método tradicional (T) y el método moderno de enseñanza basado en problemas(M). Una muestra aleatoria de 9 calificaciones finales con el método T y otra muestra aleatoria de 10 calificaciones finales con el método M dieron los siguientes resultados:

Muestra T: 06, 14, 08, 11, 10, 18, 15, 20, 13

Muestra M: 12, 11, 12, 10, 14, 15, 10, 13, 14, 12

Se asume que las calificaciones finales son dos poblaciones independientes con distribución normal.

Con un nivel de significación de 0.05, ¿es la calificación promedio del método tradicional igual a la calificación promedio del método moderno?. **Desarrolle una prueba unilateral adecuada de las dos medias.**

SOLUCION.

Sean X_1 y X_2 las variables aleatorias que representan las calificaciones finales de los métodos de enseñanza T y M respectivamente.

Se asume que $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$.

Antes de realizar la prueba de las dos medias se debe analizar la homogeneidad de varianzas.

a) Prueba de la homogeneidad de varianzas

1. *Hipótesis*: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Poblaciones normales. Suponiendo verdadera la hipótesis nula H_0, para $n_1 = 9$ y $n_2 = 10$, la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

que se distribuye como $F(8, 9)$.

4. *Región crítica*. Para $\alpha = 0.05$ y una prueba bilateral, en la distribución de $F(9, 8)$ se encuentra:

$$f_{0.975,8,9} = 4.10, \quad \text{y} \quad f_{0.025,8,9} = 1/f_{0.975,9,8} = 1/4.36 = 0.229$$

Luego, la región de rechazo de H_0 está dada por:

$$R.C. = \{F < 0.229 \quad \text{o} \quad F > 4.10\}.$$

5. *Cálculos*. De los datos de la muestra se obtiene:

$$\hat{s}_1^2 = 20.693, \quad \hat{s}_2^2 = 2.9$$

$$f_k = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{20.693}{2.9} = 7.136.$$

6. *Decisión*. Como $f_k = 7.136 \notin R.C.$ se debería rechazar H_0 con un riesgo tipo I de 0.05 y concluir que las varianzas de los métodos T y M son diferentes.

Observe que La probabilidad P es $P=P[F>7.94]=0.0039$. Luego, 0.0039 es la significación para una prueba unilateral y $2\times0.0039=0.0078$ para una prueba bilateral. Por lo que se infiere que las varianzas poblacionales son diferentes con un nivel de significación mayor a 0.0078.

b) Prueba de diferencia de las dos medias:

Como la media de la muestra de T es mayor que la media de la muestra de M es conveniente utilizar una prueba unilateral de cola derecha

1. *Hipótesis*: $H_0 : \mu_1 = \mu_2$ (Los promedios de T y M son iguales)

   $H_0 : \mu_1 > \mu_2$ (El método tradicional T tiene mejor promedio)

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística de la prueba*: Si se supone H_0 verdadera y dado que hay prueba de que las varianzas poblacionales son diferentes, la estadística apropiada es:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}},$$

que se distribuye según una t-Student con r grados de libertad, donde:

$$r = \frac{\left[\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2\right]^2}{\frac{\left[\hat{s}_1^2/n_1\right]^2}{n_1 - 1} + \frac{\left[\hat{s}_2^2/n_2\right]^2}{n_2 - 1}} = \frac{[(20.693/9) + (2.9/10)]^2}{\frac{[20.693/9]^2}{8} + \frac{[2.9/10]^2}{9}} = 10.004 \cong 10.$$

4. *Región crítica*: Para $\alpha = 0.05$ y una prueba unilateral de cola a la derecha, en la distribución $t(10)$, se encuentra $t_{0.95,10} = 1.812$. La región de rechazo de H_0 en la variación de T es: $R.C. = \{T > 1.812\}$

5. *Cálculos*: De los datos de las muestras se tiene:

$$n_1 = 9, \quad \bar{x}_1 = 12.778, \quad \hat{s}_1 = 4.549,$$
$$n_2 = 10, \quad \bar{x}_2 = 12.3, \quad \hat{s}_2 = 1.703,$$
$$\bar{x}_1 - \bar{x}_2 = 0.478$$

$$ES = \text{Error estándar de la diferencia} = \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = 1.6091$$

$$t_k = \frac{\bar{x}_1 - \bar{x}_2}{ES} = \frac{0.478}{1.6091} = 0.297$$

6. *Decisión*: Ya que $t_k = 0.297 \notin R.C.$, debemos aceptar H_0 al nivel 5% y concluir que son iguales las medias de los dos métodos. .

Observe que la probabilidad P es $P = P[T > 0.297] = 0.3863$. Luego, 0.3863 es la significación para una prueba unilateral. Por lo tanto las dos medias poblacionales son iguales a cualquier nivel menor de 0.3863.

NOTA . (Métodos Gráficos para comparar medias)

Los diagramas de caja se pueden utilizar también para comparar medias de dos o más poblaciones independientes o correlacionadas.

Como se recordará (Cap3 de la primera parte de este libro) el **diagrama de caja** ("**box plots**") contiene la mediana, los cuartiles y los valores extremos.

La caja representa el **rango intercuartil** (d) que encierra el 50% de los valores y tiene la **mediana** (o la media) dibujada dentro. Las extensiones pueden contener los valores extremos de los datos o los valores $P_{25} - 1.5d$ o $P_{75} + 1.5d$

El rango intercuartil d es la diferencia del percentil 75, P_{75} (cuartil superior) y el percentil 25, P_{25} (cuartil inferior).

Si las dos cajas no se intersectan o no se traslapan (figura 10.13a) se debería concluir que hay una diferencia significativa entre las dos medias poblacionales.

Si las dos cajas se intersectan o se traslapan (figura 10.13b), no existen reglas para concluir en forma precisa cuando hay una diferencia significativa entre dos medias poblacionales. Sin embargo, si la línea de la mayor mediana de las dos muestras excede la línea del percentil P_{75} de la otra muestra, existe evidencia de una diferencia significativa entre las dos medias a un nivel de significación que se deberá determinar utilizando la distribución de probabilidades de la estadística de la prueba.

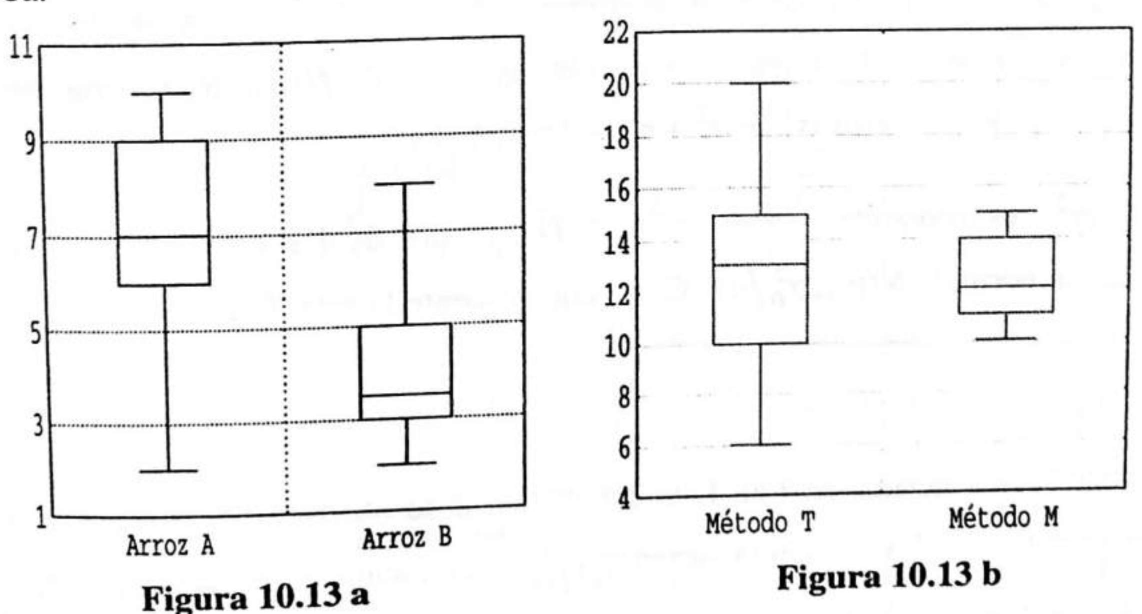


Figura 10.13 a

Figura 10.13 b

El lector debe tener en cuenta que el uso de los diagramas de caja en prueba de hipótesis no sirven como un reemplazo del procedimiento de la prueba.

Los diagramas de caja de la figura 10.13a corresponden a los datos del **ejemplo 10.10**. Las cajas no se intersectan, por lo que se concluye que las dos medias poblacionales son diferentes a cualquier nivel de significación de dos decimales.

También, los diagramas de caja de la figura 10.13b corresponden a los datos del **ejemplo 10.11**. Las cajas se intersectan, pero la mayor mediana no supera al percentil 75 de la otra caja, por lo que se concluye las dos medias poblacionales son iguales.

10.7 Prueba de la diferencia entre dos medias con observaciones aparejadas

Sea (X_1,Y_1), (X_2,Y_2),..., (X_n,Y_n) una muestra aleatoria de n datos en parejas, donde las muestras $X_1,X_2,...,X_n$, e $Y_1,Y_2,...,Y_n$, **correlacionadas**, son seleccionadas respectivamente de dos poblaciones normales $X \sim N(\mu_1,\sigma_1^2)$ y $Y \sim N(\mu_2,\sigma_2^2)$.

Podemos concebir estas n diferencias: $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2$,..., $D_n = X_n - Y_n$ como una muestra aleatoria seleccionada de una población de diferencias $D = X - Y$ cuya distribución es normal $N(\mu_D,\sigma_D^2)$, con media $\mu_D = \mu_1 - \mu_2$ y varianza $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\text{cov}(X,Y)$.

Si σ_D^2 es conocida, la estadística $\overline{D}$, media de las diferencias, tiene distribución normal $N(\mu_D,\sigma_D^2/n)$. Consecuentemente la estadística:

$$Z = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sigma_D/\sqrt{n}}$$

tiene distribución normal $N(0,1)$. Esta estadística Z se utiliza en la prueba de dos medias correlacionadas cuando la varianza σ_D^2 es conocida.

Por otro lado, si σ_D^2 *es desconocida*, entonces, la estadística,

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{\hat{S}_D/\sqrt{n}}$$

tiene distribución t-student con $n-1$ grados de libertad, en la que

$$\hat{S}_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}.$$

Esta estadística T se utiliza en la prueba de dos medias correlacionadas cuando la varianza σ_D^2 es desconocida.

Si la hipótesis $H_0: \mu_1 - \mu_2 = 0$ o su equivalente $H_0: \mu_D = 0$, se supone verdadera, entonces, la estadística de la prueba es:

$$T = \frac{\overline{D}}{\hat{S}_D / \sqrt{n}} \sim t(n-1)$$

Su valor $$t_k = \frac{\overline{d}}{\hat{s}_d / \sqrt{n}}$$

valor que resulta de dos muestras aleatorias correlacionadas, se utiliza para probar la hipótesis nula $H_0: \mu_D = 0$ contra cualquiera de las alternativas; $H_1: \mu_D \neq 0$ ó $H_1: \mu_D > 0$ ó $H_1: \mu_D < 0$.

La desviación estándar de la diferencia de las medias correlacionadas $\hat{s}_d$ se calcula por:

$$\hat{s}_d = \sqrt{\frac{\sum_{i=1}^{5}(d_i - \overline{d})^2}{n-1}} = \sqrt{\frac{\sum d^2 - n(\overline{d})^2}{n-1}}$$

Las regiones de rechazo de H_0 y las reglas de decisión de esta prueba t, son similares a las de la sección 10.3 usando la distribución t con $n-1$ grados de libertad.

1) Prueba bilateral o de dos colas

Si se prueba $H_0: \mu_1 = \mu_2$ contra $H_1: \mu_1 \neq \mu_2$, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{T < -t_{1-\alpha/2,\, n-1} \text{ o } T > t_{1-\alpha/2,\, n-1}\}$$

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{ T > t_{1-\alpha,\ n-1} \}$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 < \mu_2$, la región de rechazo de H_0 es el intervalo:

$$R.C. = \{ T < -t_{1-\alpha,\ n-1} \}$$

EJEMPLO 10.12.

Un administrador está probando la posibilidad de usar un nuevo paquete de computación. Cambiará de paquete si hay prueba que el nuevo usa menos tiempo que el antiguo al procesar determinada tarea. A fin de tomar una decisión se seleccionó una muestra aleatoria de 10 operadoras y se registró el tiempo de procesamiento en segundos con ambos paquetes tal como se da en la tabla que sigue.

Operadora	1	2	3	4	5	6	7	8	9	10
P. Antiguo	7	8	10	12	13	13	14	14	15	16
P. Nuevo	6	9	7	11	10	11	15	12	13	12
Diferenc d_i	1	−1	3	1	3	2	−1	2	2	4
d_i^2	1	1	9	1	9	4	1	4	4	16

¿Se cambiará el paquete de computo antiguo por el nuevo?.
Use un nivel de significación del 5% y haga las suposiciones necesarias.

SOLUCION.

1. *Hipótesis*: $\mu_D = \mu_1 - \mu_2$, $H_0 : \mu_D = 0$ contra $H_1 : \mu_D > 0$ (se cambiará)
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística de la prueba*: Se suponen poblaciones normales. Si la hipótesis nula es verdadera, entonces, para $n = 7$, la estadística es
$$T = \frac{\overline{D}}{\hat{S}_D / \sqrt{n}}$$
que se distribuye según una t-Student con $n - 1 = 6$ grados de libertad.
4. *Región crítica*: Para $\alpha = 0.05$ y la alternativa unilateral cola derecha en la distribución $t(9)$ se halla el valor crítico $t_{0.975,\ 9} = 2.262$. La región de rechazo de H_0 en la variación de T es el intervalo: $R.C. = \{T > 2.262\}$

5. *Cálculos*: De los datos se tiene: $n = 10$, $\bar{d} = \dfrac{16}{10} = 1.6$,

$$\hat{s}_d = \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{50 - 10(1.6)^2}{9}} = 1.64655,$$

$$ES = \text{error estándar: } \hat{s}_d / \sqrt{n} = 0.52068$$

$$t_k = \frac{\bar{d}}{\hat{s}_d / \sqrt{n}} = \frac{1.6}{0.52068} = 3.073$$

6. *Decisión*: $t_k = 3.073 \in R.C.$, deberíamos rechazar H_0 al nivel de significación dado del 5%. Por lo tanto, los tiempos promedios verdaderos no son iguales..

   Observe que la probabilidad P de la prueba es: $P = P[t_9 > 3.07289] = 0.00665$. Luego, la significación para esta prueba bilateral es $2 \times 0.00665 = 0.0133$.

   Además, dado que $\mu_1 - \mu_2 = \mu_D > 0$ el menor tiempo promedio es del paquete nuevo, se recomienda, cambiar el paquete de computo antiguo por el nuevo .

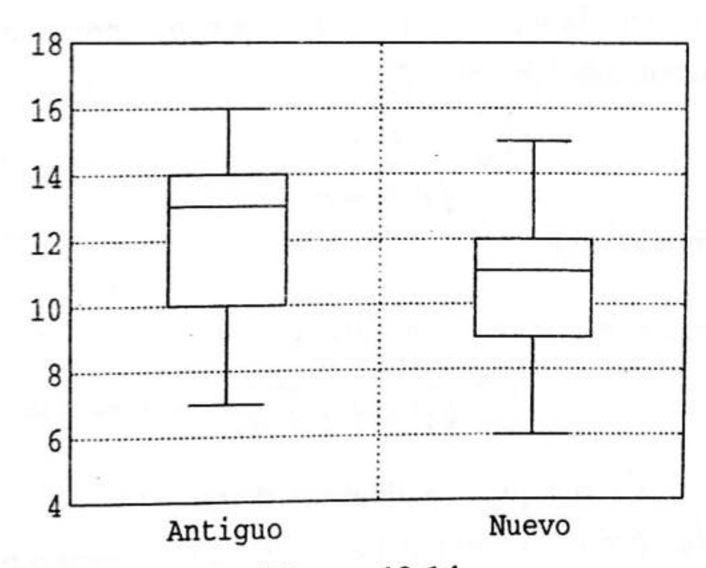


Figura 10.14

NOTA. Los diagramas de caja de la figura 10.14 corresponden a los datos del **ejemplo 10.12**. Las cajas se intersectan, pero la mayor mediana supera al percentil 75 de la otra caja, por lo que se concluye que existe diferencia significativa entre dos medias a un nivel que debe determinarse.

10.8 Prueba de hipótesis acerca de proporciones

10.8.1 Una sola proporción

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n escogida de una población Bernoullí $B(1,p)$, donde el parámetro desconocido p es la proporción de éxitos en la población, y sea la estadística,

$$\overline{P} = \frac{X}{n}$$

la *proporción de éxitos en la muestra*, siendo $X = \sum_{i=1}^{n} X_i$ el número de éxitos en la muestra una variable aleatoria binomial $B(n, p)$.

La variable aleatoria P tiene media igual a: p, varianza igual a: $np(1-p)$ y si n es suficientemente grande $(n \geq 30)$, la estadística

$$Z = \frac{\overline{P} - p}{\sqrt{p(1-p)/n}}$$

tiene distribución aproximadamente normal $N(0,1)$.

Si se supone verdadera la hipótesis nula $H_0 : p = p_0$, entonces, la distribución muestral de X es exactamente binomial $B(n, p_0)$, y la de la variable aleatoria

$$Z = \frac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

es aproximadamente normal $N(0,1)$.

La estadística

$$z_k = \frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

valor calculado a partir de una muestra aleatoria de tamaño n donde $\overline{p} = x/n$, se utiliza para probar $H_0 : p = p_0$, contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión de esta prueba Z son similares a los de la sección 10.2

1) **Prueba bilateral.** Si la prueba es de $H_0 : p = p_0$ contra $H_1 : p \neq p_0$, la región crítica o de rechazo de H_0 en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha/2} \text{ o } Z > z_{1-\alpha/2}\}$$

2) **Prueba unilateral cola derecha.** Si la prueba es de $H_0 : p = p_0$ contra $H_1 : p > p_0$, la región crítica o de rechazo de H_0 en los valores de Z es el intervalo:

$$R.C. = \{Z > z_{1-\alpha}\}$$

3) **Prueba unilateral cola izquierda.** Si la prueba es de $H_0 : p = p_0$ contra $H_1 : p < p_0$, la región crítica o de rechazo de H_0 en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha}\}$$

NOTA. Al sustituir la variable $Z = \dfrac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}}$ en *R.C.* se obtienen las regiones críticas respectivas en los valores de las variables $\overline{P}$ y X respectivamente.

NOTA. (Muestras pequeñas) Sea x la cantidad de éxitos en una muestra aleatoria pequeña de tamaño n (n<30)
Prueba bilateral. Si $x < np_0$ se calcula

$$P = P[X \leq x \text{ cuando } p = p_0] = \sum_{k=0}^{x} C_k^n \, p_0^k (1-p_0)^{n-k}$$

y si $x > np_0$ se calcula

$$P = P[X \geq x \text{ cuando } p = p_0] = \sum_{k=x}^{n} C_k^n \, p_0^k (1-p_0)^{n-k}.$$

Se rechaza $H_0 : p = p_0$ si $P \leq \alpha/2$. En caso contrario no se rechaza H_0.

Prueba unilateral cola derecha . Se calcula

$$P = P[X \geq x \text{ cuando } p = p_0] = \sum_{k=x}^{n} C_k^n \, p_0^k (1-p_0)^{n-k}$$

y se rechaza $H_0 : p = p_0$ si el valor de P es menor o igual que el nivel de significación α.

Prueba unilateral cola izquierda. Se calcula

$$P = P[X \leq x \text{ cuando } p = p_0] = \sum_{k=0}^{x} C_k^n \, p_0^k (1-p_0)^{n-k}$$

y se rechaza $H_0 : p = p_0$, si el valor de P es menor o igual que el nivel de significación α.

EJEMPLO 10.13.

El gerente de un banco afirma que el porcentaje de clientes que son atendidas en las ventanillas por más de 5 minutos es igual a 0.30. Con el fin de evaluar esta afirmación se escogió una muestra aleatoria de 400 clientes atendidos en las ventanillas del banco y se encontró que 100 de ellos demoraron más de 5 minutos.

a) Al nivel de significación del 1%, ¿presenta esta muestra suficiente evidencia que indique que el porcentaje de clientes que demoran más de 5 minutos en las ventanillas es diferente de 0.30?.

b) Calcular la probabilidad de tomar la decisión errada de aceptar la afirmación del gerente cuando la verdadera proporción de todos los clientes que sobrepasan los 5 minutos de atención es 0.20.

SOLUCION.

a) Sea p el porcentaje poblacional de atención que supera los 5 minutos.

1. *Hipótesis*: $H_0 : p = 0.30$ contra $H_1 : p \neq 0.30$
2. *Nivel de significación* $\alpha = 0.01$.
3. *Estadística*: Si $H_0 : p = 0.30$, es supuesta verdadera, y n grande la estadística es:

$$Z = \frac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\overline{P} - 0.3}{\sqrt{0.3(0.7)/n}}$$

que tiene distribución aproximadamente normal $N(0,1)$.

4. *Región crítica*: Para $\alpha = 0.01$ y una alternativa bilateral, en la distribución de Z se encuentra el valor crítico $z_{0.995} = 2.575$. Luego, la región crítica o de rechazo de H_0 en Z es el intervalo:

$$R.C. = \{Z < -2.575 \ \text{ o } \ Z > 2.575\}.$$

5. *Cálculo*: $n = 400,\ x = 100,\quad \overline{p} = \frac{x}{n} = \frac{100}{400} = 0.25$

$$ES = \text{Error estándar} = \sqrt{p_0(1-p_0)/n} = \sqrt{0.3 \times 0.7/400} = 0.0229$$

$$z_k = \frac{\overline{p} - p_0}{ES} = \frac{0.25 - 0.3}{0.0229} = -2.18$$

6. *Decisión*: Como $z_k = -2.18 \notin R.C.$, no deberíamos rechazar H_0 al 5%.

Observe que la significación bilateral es $2\times P = 2\times P[Z>2.18] = 2\times 0.0146 = 0.0292$.

b) Al sustituir en *R.C.* la variable: $Z = \dfrac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{\overline{P} - 0.3}{0.0229}$,

se obtiene la región crítica de rechazo de H_0 en la variable $\overline{P}$:

$$R.C. = \{\overline{P} < 0.241 \text{ o } \overline{P} > 0.359\}$$

Luego, la probabilidad de aceptar $H_0 : p = 0.30$ cuando realmente $p = 0.20$ (error tipo II) es igual a:

$$\beta = P[\text{aceptar } H_0 / p = 0.20] = P[0.241 \le \overline{P} \le 0.359 / p = 0.20]$$

$$\beta = P\left[\frac{0.241-0.20}{0.02} \le \frac{\overline{p}-0.20}{\sqrt{(0.2\times 0.8)/400}} \le \frac{0.359-0.20}{0.02}\right]$$

$$\beta = P[2.05 \le Z \le 7.95] = 0.0202$$

EJEMPLO 10.14

Se afirma que cierto medicamento que se prescribe para aliviar determinada enfermedad es efectivo en más del 80% de los casos. Al parecer esta afirmación es exagerada por lo que se suministró tal medicamento a una muestra aleatoria de 15 pacientes resultando que 13 de ellos han experimentado alivio, ¿es ésta suficiente evidencia para concluir que realmente el medicamento es efectivo en más del 80% de los casos al nivel de significación del 5%?.

SOLUCION.

Sea X el número de pacientes que se sanan en $n = 15$ casos. Entonces $X \sim B(15, p)$ donde p es el porcentaje de pacientes que se sanan en la población de todos los pacientes que sufren la enfermedad.

1. *Hipótesis*: $H_0 : p = 0.80$ contra $H_1 : p > 0.80$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Si la hipótesis nula es cierta, la variable X tiene distribución binomial con $n = 15$ y $p = 0.8$.
4. *Región crítica*: Se rechazará H_0 en favor de H_1 si el valor de $P = P[X \ge 13 \text{ cuando } p = 0.80]$ es menor que $\alpha = 0.05$
5. *Cálculo*:

$$P = P[X \ge 13/p = 0.80] = \sum_{k=13}^{15} C_k^{15} (0.8)^k (0.2)^{15-k} = 0.3970.$$

6. *Decisión*: Dado que $P = 0.3970 > \alpha = 0.05$, no se debe rechazar H_0.

10.8.2 Dos proporciones con observaciones independientes

Sean X_1 y X_2 el número de éxitos en dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones de Bernoullí $B(1, p_1)$ y $B(1, p_2)$, donde los parámetros desconocidos p_1 y p_2 son las proporciones de éxitos poblacionales respectivos.

Sean además las proporciones de éxitos muestrales respectivas:

$$\overline{P}_1 = \frac{X_1}{n_1} \quad \text{y} \quad \overline{P}_2 = \frac{X_2}{n_2}$$

Para n_1 y n_2 suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$), la variable aleatoria

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}},$$

tiene distribución aproximadamente normal $N(0,1)$.

Si $H_0 : p_1 = p_2$ se supone verdadera, la estadística es;

$$Z = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\dfrac{p_c(1-p_c)}{n_1} + \dfrac{p_c(1-p_c)}{n_2}}} \cong N(0,1)$$

donde, p_c **es el valor común de los parámetros** p_1 y p_2 **cuya estimación insesgada** (probar!) es:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \overline{p}_1 + n_2 \overline{p}_2}{n_1 + n_2}$$

La estadística

$$z_k = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_1} + \dfrac{\hat{p}(1-\hat{p})}{n_2}}}.$$

valor que resulta de dos muestras aleatorias, se utiliza para probar la hipótesis nula $H_0: p_1 = p_2$, contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión de esta prueba Z son similares a los de la sección 10.2

1) **Prueba bilateral.** Si la prueba es de $H_0: p_1 = p_2$ contra $H_1: p_1 \neq p_2$, la región crítica o de rechazo de H_0 en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha/2} \quad \text{o} \quad Z > z_{1-\alpha/2}\}$$

2) **Prueba unilateral cola derecha.** Si la prueba es de $H_0: p_1 = p_2$ contra $H_1: p_1 > p_2$, la región crítica o de rechazo de H_0 en los valores de Z es el intervalo:

$$R.C. = \{Z > z_{1-\alpha}\}$$

3) **Prueba unilateral cola izquierda.** Si la prueba es de $H_0: p_1 = p_2$ contra $H_1: p_1 < p_2$, la región crítica o de rechazo de H_0 en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha}\}$$

EJEMPLO 10.15.

Un patrocinador de un programa especial de televisión afirma que el programa representa un atractivo mayor para los televidentes hombres que para las mujeres, pero, el personal de producción del programa piensa que es igual el porcentaje de televidentes hombres y mujeres que ven el programa especial. Si una muestra aleatoria de 300 hombres y otra de 400 mujeres reveló que 120 hombres y 120 mujeres estaban viendo el programa especial de televisión, ¿Puede considerarse significativa la diferencia al nivel $\alpha = 5\%$?

SOLUCION.

Sean p_1 y p_2, respectivamente, las proporciones de hombres y mujeres que ven el programa especial de televisión.

1. Hipótesis $H_0: p_1 = p_2$ contra $H_1: p_1 > p_2$.
2. Nivel de significación: $\alpha = 0.05$.
3. Estadística. Si $H_0: p_1 = p_2$ es supuesta verdadera y las muestras son grandes, la estadística es:

$$Z = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}$$

que tiene distribución aproximadamente normal $N(0,1)$.

4. Región crítica. Para $\alpha = 0.05$ y una prueba unilateral cola a la derecha, la región crítica o de rechazo de H_0 es:

$$R.C. = \{Z > 1.645\}$$

5. Cálculo. Los datos de la muestra dan

Hombres	mujeres
$n_1 = 300$	$n_2 = 400$
$p_1 = 120$	$p_2 = 120$

$$\overline{p}_1 = \frac{120}{300} = 0.4\,, \qquad \overline{p}_2 = \frac{120}{400} = 0.3$$

$$\hat{p} = \frac{120+120}{300+400} = 0.34$$

$$ES = \text{Error estándar} = \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} = \sqrt{\frac{0.34 \times 0.66}{300} + \frac{0.34 \times 0.66}{400}} = 0.03618$$

$$z_k = \frac{\overline{p}_1 - \overline{p}_2}{ES} = \frac{0.4 - 0.3}{0.03618} = 2.764$$

6. Decisión: Como $z_k = 2.764 \in R.C.$, deberíamos rechazar H_0.

EJERCICIOS

Error tipo I y error tipo II

1. Se seleccionó una muestra aleatoria de tamaño 64 de una población con media μ y varianza 40^2 para realizar la prueba de la hipótesis: $H_0 : \mu = 150$, contra $H_1 : \mu < 150$.
   a) Describa la regla de decisión en la estadística Z, si la probabilidad de error tipo I es 0.015.
   b) Si la media de la muestra resulta 141, ¿cuál es la decisión respecto a H_0?
   c) Halle el valor de la probabilidad P. Interprete el significado

   Rp. a) $RC=\{Z<-2.17\}$ $\overline{X} < 139.15$ }, b) $Z_{calc}=-1.8$, se acepta H_0, c) 0.0359

2. Se seleccionó una muestra aleatoria de tamaño 36 de una población con media μ para realizar la prueba de la hipótesis: H_0: $\mu=100$, contra H_1: $\mu \neq 100$. De tal muestra se obtuvo la media 108 y la desviación estándar 24.
   a) Describa la regla de decisión en $\overline{X}$ si el nivel de significación es 0.05.
   b) ¿Cuál es la decisión con respecto a H_0, en el nivel dado?
   c) Halle el valor de P. De su comentario sobre este valor

   Rp. a) $RC=\{ \overline{X} < 92.16$ o $\overline{X} > 107.84\}$, b) se rechaza H_0, c) 2×0.228

3. Se seleccionó una muestra aleatoria de 16 observaciones de una población normal con media μ y varianza 400. Si para realizar la prueba de la hipótesis nula: $H_0 : \mu = 60$ se utiliza la región de rechazo $RC = \{ \overline{X} > 70.6\}$.
   a) Determine la probabilidad de error tipo I
   b) Determine la probabilidad de error tipo II, si realmente, $\mu = 75$.
   c) Ilustre en una gráfica la solución del problema

   Rp. a) $\alpha = 0.017$, b) $\beta = 0.1894$

4. Se seleccionó una muestra aleatoria de 36 observaciones de una población con media μ desviación estándar 18. Si para realizar la prueba de la hipótesis nula: $H_0 : \mu = 50$, se utiliza la región de rechazo: $RC = \{ \overline{X} < 43$ o $\overline{X} > 57\}$.
   a) Determine el nivel de significación de la prueba
   b) Determine la probabilidad de error tipo II, si realmente, $\mu=41$.

   Rp. a) $\alpha = 0.0456$, b) $\beta = 0.0668$

5. Una firma va a comercializar un nuevo producto sólo si hay prueba de que al menos el 20% de todos los consumidores lo prefieren. Para probar esa hipótesis se va a seleccionar al azar 400 consumidores. Si se utiliza como región crítica $\{X < 60\}$ donde X es el número de consumidores en la muestra que prefieren el producto, calcule:
   a) El nivel de significación
   b) La probabilidad de cometer error tipo II si realmente p=0.10

Rp. a) $\alpha=P[X<60/p=0.2]=P[Z<-2.5]=0.0062$,
b) $\beta=P[X\geq 150/p=0.10]=P[Z\geq 3.33]=0.0004$

6. Una familia cosiste de 10 personas en las que k de ellas sufren alguna enfermedad y el resto son sanas. Para realizar la prueba de $H_0: k = 4$ contra $H_1: k < 4$, se seleccionan de esa familia 3 personas al azar. Si dos de las tres sufren alguna enfermedad se rechaza H_0, en caso contrario se acepta H_0.
   a) Halle la probabilidad de error tipo I
   b) Halle la probabilidad de error tipo II si $k = 3$.

Rp. a) $\alpha=P$[1 sano 2 enfermos / k=4]=0.114, b) β=P[NO (1 sano 2 enfermos) / k=3]=0.967.

7. Un sistema automático de producción está controlado si el 90% de la producción no es defectuosa. Para comprobar esta hipótesis se seleccionan al azar 10 objetos de la producción y se decidirá rechazar que el proceso está controlado si menos de ocho objetos no son defectuosos. Suponiendo independencia:
   a) Plantee las hipótesis de esta prueba
   b) Obtenga el nivel de significación
   c) Determine la potencia del contraste si sólo 60% de la producción es no defectuosa.

Rp. a) H_0: p=0.9, H_1: p<0.9, X::# objetos buenos b) $\alpha=P[X<8/\ p=0.9]=0.0702$, $X\sim B(10, 0.9)$,
c) $1-\beta=P[X<8/\ p=0.6]=0.8327$, $X\sim B(10, 0.6)$

8. En una cuadra viven 100 familias de los cuales r tienen al menos una tarjeta de crédito y el resto no. Para docimar la hipótesis nula H_0: $r = 30$ contra H_1: $r > 30$ se seleccionan al azar a 20 familias. Si en la muestra se encuentran por lo menos 10 familias con por lo menos una tarjeta de crédito se decidirá rechazar H_0. Calcule la probabilidad de cometer error tipo II cuando $r = 40$.

Rp. $X\sim H(100, m, 20)$, $\beta=P$[aceptar H_0/m=40]= $1-P[X<10/m=40]=1-\sum_{k=0}^{9} C_k^{40} C_{20-k}^{60} / C_{20}^{100}$.

9. Se ha determinado que el tiempo en horas de operación de un sistema entre una falla y la siguiente tiene distribución exponencial con parámetro β. Para probar la hipótesis nula H_0: β = 1/10 cada cierto tiempo se hace **una** medición del

tiempo X entre dos fallas consecutivas y se decide que si $X < 9$ se rechazará H_0, de otro modo se aceptará H_0.

a) Calcule el nivel de significación de la prueba
b) Determine la probabilidad de error tipo II cuando $\beta = 1/8$

Rp. a) H_0: $\mu=10$, H_1: $\mu=8$, $RC=\{X<9\}$ $\alpha=P[X<9 / \mu=10]=1-e^{-9/10} \cong 0.593$,
b) $\beta=P[X \geq 9 / \mu=8]=e^{-9/8} \cong 0.325$

10. El tiempo en horas de diagnóstico de motores de automóviles se asume que es una variable aleatoria continua X cuya función de densidad es:

$$f(x)=\begin{cases}\theta x & \text{si } 0 \leq x \leq \sqrt{\dfrac{2}{\theta}} \\ 0 & \text{en otro caso}\end{cases}$$

en donde $\theta > 0$ es un parámetro desconocido.

Para realizar la prueba de la hipótesis nula H_0: $\theta = 1/2$ contra H_1: $\theta < 1/2$ se mandará a reparar uno de estos motores, si en tal reparación se demoran C horas o más se aceptará la hipótesis nula.

Determine el valor de C si se quiere una probabilidad de 5% de rechazar la hipótesis nula cuando es realmente verdadera.

Rp. $0.05.=P[\text{Rechazar } H_0 / \theta=1/2]=P[X<C/\theta=1/2]=C^2/4$, $C=0.447$.

11. El tiempo de vida útil en meses de cierto tipo de resistencia eléctrica es una variable aleatoria X que tiene distribución exponencial con media μ. Para comprobar la hipótesis $H_0: \mu = 80$ contra, $H_1: \mu < 80$, se usa como región de rechazo de H_0 el intervalo $\{X<77.5\}$ donde X es la vida útil de una muestra aleatoria de tamaño 1 escogida de la población de resistencias.

a) Halle la probabilidad de error tipo I.
b) Halle la potencia de la prueba cuando la verdadera media es 76 meses.

a) Rp. H_0: $\mu=80$, H_1: $\mu < 80$, $R.C.=\{X < 77.5\}$, $\alpha=P[X < 77.5 / \mu=80] = 1-e^{-77.5/80} \cong 0.620$, b)
$\beta=P[X \geq 77.5 / \mu=76] = e^{-9/8} \cong 0.3607$, $1-\beta=0.6397$.

12. El tiempo de falla, en horas, de cierta componente electrónica es una variable aleatoria X que tiene distribución exponencial con parámetro β. Para realizar la prueba de la hipótesis nula $H_0: \beta = 0.01$ contra $H_1: \beta < 0.01$ se escogen dos de tales componentes al azar y se observa sus tiempos de falla. Si se decide rechazar H_0 cuando el tiempo de falla de por lo menos una de ellas es menor que 400 horas, halle la probabilidad de error tipo I.

Rp. $\alpha=P[(X_1<400 \text{ o } X_2<400)/ \beta=0.01]=P[(X_1<400)/\beta=0.01]+ P[(X_2<400)/\beta=0.01]-$
$P[(X_1<400 \wedge X_2<400)/\beta=0.01]=2p-p^2$, donde $p=P[(X_1<400)/\beta=0.01] = 1-e^{-400/100} \cong 0.982$.

Una media

13. El gerente de la cadena de tiendas REPLAY afirma que en promedio cada cliente gastó \$500 el año pasado. Sin embargo analizando el mercado, nosotros creemos, que dicho gerente ha exagerado. Para someter a prueba estas hipótesis se tomó una muestra aleatoria de 100 clientes que el año pasado habían comprado en dicha tienda, ésta reveló una media de \$470 y una desviación estándar \$100.

    a) En el nivel de significación $\alpha = 0.015$, ¿es posible concluir que los clientes de esta tienda están gastando menos?

    b) Si la verdadera media del consumo es 450\$, halle la probabilidad de aceptar la hipótesis del gerente.

    Rp. a) H_0: $\mu=500$, H_1:$\mu<500$, $RC=\{\overline{X}<478.3\}$, se rechaza H_0
    b) $\beta=P[\overline{X}\geq 478.3]=P[Z\geq 2.83]=0.0023$.

14. El gerente de ventas de la empresa "Gato S.A" que elabora cápsulas de uña de gato indica que la demanda semanal tiene distribución normal con una media de 1000 cápsulas y una desviación estándar de 360 cápsulas. Sin embargo en un estudio reciente una muestra aleatoria de 36 semanas dio una demanda promedio de 850 cápsulas.

    a) En el nivel de significación de 0.05, ¿es posible concluir que la media de la producción semanal es menos de 1000 cápsulas?.

    b) Determine la probabilidad P de la prueba

    Rp. a) $H_0: \mu \geq 1{,}000$, $H_1: \mu < 1{,}000$, $z_k=-2.5$, $RC=\{Z<-1.645\}$, se rechaza H_0 si.
    b) $P=P[Z<-2.5]=0.062$.

15. Un proceso de llenado automático de durazno en rodajas está preocupando al gerente de producción por que las latas se están llenando en exceso. Por registros anteriores se sabe el peso neto (en gramos) de las latas tiene distribución normal con media 500 y que el 95% de todos esos pesos están entre 480.4 y 519.6. El departamento de control de calidad tomó una muestra al azar de 9 latas de la producción y obtuvo los siguientes pesos:

    490, 495, 501, 492, 490, 500, 493, 502, 501

    a) En el nivel de significación $\alpha=0.05$, ¿es posible concluir que el peso medio es diferente a 500 gramos?

    b) Determine la probabilidad P de la prueba

    Rp. a) $H_0: \mu=500$, $H_1: \mu\neq 500$, $\sigma=10$, $RA=\{-1.96\leq Z\leq 1.96\}$, $z_k=-1.2$, No., b) $P=2\times 0.1151$

16. Un reporte estadístico, afirma que los fumadores adultos de cigarrillos en Lima consumen en promedio 10 cigarrillos por día. Un grupo de alumnos que hace un trabajo en estadística aplicada va a comprobar esta afirmación. Para esto ha escogido una muestra aleatoria de 36 fumadores adultos y se observó una media de 9 y una desviación estándar de 3 cigarrillos por día.
   a) En el nivel de significación de 0.05, ¿es posible concluir que el consumo promedio de cigarrillos ha bajado?.
   b) Determine la potencia de esta prueba si el valor real de la media es 8 cigarrillos por día.

   Rp. a) H_0:μ =10, H_1:μ<10, z_k=−2, RC={Z<−2.33} , se acepta H_0. b) 1 −β=1−0.0475=0.9525

17. El gerente de producción de la compañía de cerveza "DORADA" revisa su línea de producción. El llenado automático debe dar un contenido medio 320 cm^3. Una muestra aleatoria de 36 latas de cerveza de su producción ha dado un contenido medio de 317 cm^3 y una desviación estándar de 12 cm^3
   a) Determine la región de rechazo para una prueba unilateral en el nivel de significación 0.015. ¿Hay suficiente razón para creer que existe una baja en la media de los contenidos?.
   b) ¿Con qué probabilidad esta prueba **no detecta** una diferencia igual a 8 cm^3 en el promedio de los contenidos y por debajo de lo que indica la hipótesis nula?.

   Rp. H_0: μ=320, H_1:μ<320 , z_k=−1.5, RC={ $\overline{X}$ <315.56}, No.
   b) P[aceptar H_0 /μ=312]=P[Z≥1.78]=0.0375

18. La prueba de resistencia física que se aplica a los alumnos de la PUCP tiene una media de 200 puntos y una desviación estándar de 50 puntos. Para comprobar la hipótesis de la media se sometieron a la prueba a 100 alumnos seleccionados al azar. Si se utiliza como región de rechazo de H_0, el intervalo: { $\overline{X}$ <190}.
   a) Determine la probabilidad de tomar la decisión correcta de aceptar que H_0:μ=200 cuando realmente es verdadera.
   b) ¿Con qué probabilidad esta prueba **detecta** una diferencia igual a 15 puntos en el promedio de la resistencia y por debajo de lo que indica la hipótesis nula?

   Rp. a) RC={ $\overline{X}$ <190} P[aceptar H_0 /μ=200]=P[Z≥−2]=0.9772
   b) P[rechazar H_0 /μ=185]=P[Z≥1]=0.1587.

19. El gerente de producción de la empresa "HILOS" afirma que el nuevo hilo sintético que produce su compañía tiene una resistencia media a la ruptura mayor de 15 kilogramos. Usted piensa que esta cifra es exagerada y pide realizar una prueba. En una muestra aleatoria de 36 de tales hilos se ha medido la resistencia X_i resultando las siguientes sumas:

$$\sum_{i=1}^{36} X_i = 612, \quad \sum_{i=1}^{36} X_i^2 = 10719$$

a) Plantee las hipótesis del problema
b) ¿Cuál es la estadística de la prueba?
c) Realice la prueba de las hipótesis en el nivel de significación 0.05
d) Halle el porcentaje de las veces en que tal muestra nos lleva a rechazar en forma acertada que la resistencia media a la ruptura es igual a 15 Kg. cuando realmente es igual a 2 Kg. por encima de ello.

Rp. a) H_0:μ =15, H_1:μ >15, b) Z, c) RC={ $\overline{X}$ >15.8225}, se rechaza H_0, d) 1−β=0.9909 en 9909 casos de 10,000.

20. Un estudio estadístico indica que el tiempo en minutos que utilizan los 100 operarios para confeccionar un pantalón de la firma "JEAN" es una variable aleatoria cuya distribución es normal con media 15 y desviación estándar 3.2. Para comprobar el tiempo promedio se escogieron los tiempos de producción de 16 operarios resultando una media de 16.
a) Plantee las hipótesis adecuadas del problema.
b) ¿Cuál es la estadística de la prueba?
c) Realice la prueba de las hipótesis en el nivel de significación 0.05

Rp. a) H_0:μ =15, H_1:μ >15, b) Estd. Z, población FINITA c) z_k=1.36. RC={Z>1.645} se acepta H_0.

21. La empresa agroindustrial "COCONA S.A." de Iquitos procesa palmito y las envasa en frascos para su consumo. Se sabe que el tiempo del proceso tiene distribución normal con una media de 10 minutos. Se introduce un nuevo método para reducir el tiempo medio del proceso. Para comprobar el cambio del promedio se observaron aleatoriamente 10 tiempos de proceso y se obtuvieron los siguientes resultados.

Tiempos	7	8	9	10	11	12	12.5
# de frascos	2	2	2	1	1	1	1

a) En el nivel de significación del 0.05, ¿se puede decir que el tiempo de llenado con el nuevo método es significativamente inferior al anterior?.
b) Determine la probabilidad P de la prueba.

Rp. H_0:μ=10, H_1:μ<10, ES=0.6238, t_k=−1.0419, gl=9, a) RC={T<−1.833}, se acepta H_0. b) $MCEST$ da signific. Unilateral: P=P[T<−1.0419]=0.1623

22. Las cajas de avena llenadas por un proceso automático deben tener un contenido de 160 gramos en promedio. Si no es así debe detenerse la producción para regular la máquina. Para el control se obtuvo el peso en gramos X_i de 10 cajas seleccionadas al azar de esa producción y resultaron las siguientes sumas:

$$\sum_{i=1}^{10} X_i = 1580, \quad \sum_{i=1}^{10} X_i^2 = 249658$$

a) Plantee las hipótesis para una prueba unilateral cola izquierda.

b) ¿Qué estadística se debería usar en la prueba y que condición fundamental requiere ésta estadística?

c) En el nivel de significación de 0.01, ¿es razonable detener la producción?.

Rp. a) H_0:μ=160, H_1:μ<160, b) *t*-student, población normal, $\bar{x} = 158$, $\hat{s} = 2$, ES=0.63246, t_k=−3.16., RC={T<−2.821}, se rechaza H_0, Si.

23. Una máquina produce cierta parte componente cuya longitud debería ser 1.2 cm. promedio. Por un estudio anterior se sabe que la longitud de la componente se distribuye según la ley de probabilidad normal con desviación estándar de 0.5 cm. Existe la preocupación de que han cambiado los ajustes realizados a la máquina que las produce. Para salir de la preocupación se requiere diseñar una prueba de hipótesis con probabilidad de error tipò I igual a 0.0287.

a) Halle el tamaño de la muestra y la región crítica sabiendo que si la verdadera media es 1.6 cm, entonces, la probabilidad de error tipo II sería igual a 0.0179.

b) Si con el tamaño de la muestra hallado en a) resulta que Z = −2, halle la media de la muestra, ¿cuál sería su opinión al respecto?

Rp. a) H_0:μ=1.2, α=0.0287, da Z=1.9., β=0.0179 da Z=−2.1, y $n = 25$, RC={ $\bar{X}$ > 1.39}, b) Z=(Media−1.2)/0.1=−2, Media=1, se acepta H_0

24. Se sabe que las ventas diarias de la compañía "P&C" tienen distribución normal con media de $2277 y desviación estándar de $300. El gerente de ventas de la compañía cree que la media de las ventas ha bajado a $1800. Diseñe una prueba para estas hipótesis considerando la del gerente como una alternativa de manera que haya una probabilidad igual a 0.004 de cometer error tipo I y una probabilidad de error tipo II igual a 0.017. Ilustre con una gráfica.

Rp. $n = 9$, RC={ $\bar{X}$ < 2012}

25. En un estudio estadístico del año pasado se afirma que los ingresos familiares mensuales en el distrito de San Isidro tienen una media $400 y una desviación estándar de $100. Diseñe una prueba para probar la hipótesis de la media con un riesgo de 0.17% de rechazarla si es verdadera y con un riesgo de aceptarla en 572 de 10,000 casos cuando la media es realmente $363. Ilustre gráficamente.

Rp. 0.017=P[$\bar{X}$ <K / μ=400], implica (K−400)/(100/(n)$^{1/2}$ =−2.12, 0.0571=P[$\bar{X}$ ≥K / μ=363], implica (K−363)/(100/(n)$^{1/2}$ =1.58, entonces, n=100, K=378.8

26. El gerente de ventas de una consorcio "BANK" argumenta que el número de llamadas mensuales de los representantes de ventas hace un promedio de 400 con una desviación estándar de 56. Diversos informes dicen que la estimación de la media es muy baja motivo por el cual se debe realizar una prueba de hipótesis. Si la media de las llamadas aumenta en 28, este cambio debe detectarse con probabilidad 0.9772. Si no hay cambio, este debe detectarse con probabilidad 0.9332. Determine el número de representantes que deben seleccionarse para la muestra y la región crítica. Ilustre con una gráfica.

Rp. a) 49, b) $RC=\{\overline{X}>162\}$.

27. Un proceso automático de llena vasos de refrescos cuyo contenido debe tener una media de 20 onzas con una desviación estándar de 0.6 onzas. En forma periódica se controlan una muestra de 36 vasos de refrescos llenados por el proceso automático y se concluye que el sistema está fuera de control si la media de los contenidos de la muestra está fuera del intervalo [19.85, 20.15].
   a) Formule la hipótesis nula y alternativa
   b) Describa en que consiste los errores tipo I y tipo II.
   c) Indique la estadística apropiada de esta prueba
   d) Calcule la probabilidad de cometer error tipo I.
   e) Calcule la potencia de la prueba cuando la verdadera media sea 20.4
   f) Si usted quiere que $\alpha=0.05$ y $\beta=0.10$, halle el tamaño de muestra requerido sabiendo que la verdadera media del proceso es 20.3.

Rp. a) $H_0: \mu=20$ y $H_1: \mu\neq 20$, c) Z, d) 0.1336, e) 0.0062, f) n=51.9841≅52.

Una proporción

28. Un informe estadístico indica que el 13% de los conductores de fin de semana conducen bajo los efectos del alcohol. Sin embargo el último fin de semana fueron intervenidos aleatoriamente 500 conductores encontrándose que 80 de ellos estaban bajo los efectos del alcohol. Se quiere comprobar el informe estadístico mediante una prueba de hipótesis bilateral
   a) Plantee las hipótesis de la prueba
   b) En el nivel de significación α=0.03, determine la región de rechazo de la hipótesis nula
   c) Con la regla de decisión que resulta de b), ¿cuál es su decisión con respecto a este informe estadístico?
   d) Halle la probabilidad P de la prueba.

Rp. a) $H_0: p=0.13$, $H_1: p\neq 0.13$, b) $RA.=\{-2.17\leq Z\leq 2.17\}$, c) $z_k=2$, se acepta H_0, d) 0.0456.

29. Un fabricante afirma que es 5% el porcentaje de su producción defectuosa. Para comprobar esta hipótesis se seleccionó una muestra aleatoria de 40 observaciones y se encontró que el porcentaje muestral de defectuosos es 0.10. En el nivel de significación de 0.05
    a) Plantee las hipótesis adecuadas
    b) Determine la región rechazo de la hipótesis nula
    c) Calcule el valor de la estadística de la prueba
    d) ¿Cuál es su decisión respecto a la afirmación del fabricante?

Rp. a) H_0: p=0.05, H_1: p>0.05, b) RC={Z>1.645}, c) Z_k=1.45, d) se acepta H_0.

30. Un candidato político consistentemente ha sido favorecido por al menos 58% de la votación de la población en encuestas hechas durante los meses que preceden a las elecciones generales. Sin embargo, una encuesta a 500 votantes realizada la semana final de la campaña reveló que la proporción de votantes a favor de él fue de 54%. Al nivel de significación del 5%, ¿debería este candidato creer que el nivel de su apoyo ha bajado?.

Rp. $H_0 : p \geq 0.58$, $H_1 : p < 0.58$, $z_k = -1.81$, $R.C$={Z<−1.645}, se rechaza H_0, Si.

31. Un legislador desea probar la hipótesis que más del 65% de sus representados está a favor de cierta legislación laboral que se está presentando en el congreso, basándose en una muestra al azar de 400 ciudadanos.
    Si la probabilidad de error tipo I es $\alpha = 0.05$.
    a) ¿Qué valor como mínimo debe tener la proporción de la muestra, para que a partir de ese valor, la decisión sea aceptar la hipótesis del legislador?.
    b) ¿Cuál es la probabilidad de tomar la decisión errada de rechazar la propuesta del legislador cuando en realidad el 70% de los votantes acepta la legislación laboral?

Rp. a) K=0.689, b)$P[\overline{P} \leq 0.689 / p=0.7] = P[Z \leq -0.48]=0.3156$.

32. Un informe médico indica que el 80% de los adultos mayores de 50 años han sido intervenidos quirúrgicamente al menos una vez. Para realizar una prueba de hipótesis unilateral de este informe médico se toma una muestra aleatoria de 100 adultos mayores de 50 años. Si se utiliza como región de rechazo de la hipótesis nula el conjunto {X<72}, en donde X es el número de personas que han sido intervenidos quirúrgicamente al menos una vez
    a) Halle la probabilidad de error tipo I
    b) Halle la probabilidad de error tipo II, si el porcentaje real es 0.65
    c) Ilustre la solución del problema con una gráfica.

Rp. a) $\alpha=P[Z<-2]=0.0228$, b) $\beta=P[Z\geq 1.47]=0.0708$

33. La compañía de productos lácteos "La Leche" está considerando cambiar sus actuales envases plásticos por envases de cartón, siempre y cuando se compruebe que más del 60% de los consumidores aceptan el nuevo envase de cartón, para esto realiza una consulta a 600 consumidores seleccionados al azar.

a) En el nivel de significación no mayor de 0.05, ¿qué valor mínimo debe tener el número de consumidores en la muestra que aceptan el nuevo envase de cartón para que a partir de ese valor la decisión sea cambiar los envases?.

b) Si 384 consumidores de la muestra aceptan el nuevo envase, ¿cuál es la significación para esta prueba?.

Rp. a) $H_0 : p \leq 0.60$, $H_1 : p > 0.60$, $R.C. =]1.645, +\infty[$, se decide cambiar de envase si más de 380 consumidores desean el cambio., b) $P= P[Z>2]=0.0228$.

34. La empresa encuestadora C&A afirma que el 20% de todos los electores de la población están a favor del candidato a Alcalde: Sr. Ruiz. Queremos comprobar esta afirmación utilizando la siguiente regla de decisión: Si la proporción a favor del Sr. Ruiz, en una muestra al azar de 400 electores está entre 16.08% y 23.92% se decide aceptar la hipótesis de la encuestadora. En caso contrario se decide rechazar tal hipótesis.

a) Determine el nivel de significación de la prueba.

b) Halle la probabilidad cometer error tipo II si realmente el porcentaje a favor es 0.25.

Rp. a) $\alpha =0.05$, b) $\beta=0.3085$.

35. La distribuidora "PERUFAM" va a comercializar un nuevo producto sólo si se comprueba que al menos el 20% de todos los consumidores lo prefieren. Se escogió una muestra al azar de 100 consumidores y se halló el número X de consumidores que prefieren el producto.

a) Si se utiliza como región de rechazo: $\{ X < 15\}$, calcule el nivel de significación de la prueba.

b) Halle el valor crítico K si se desea rechazar la hipótesis nula cuando $X< K$ al nivel de significación no mayor de 0.05

c) ¿Con qué probabilidad la prueba detecta que la hipótesis nula es falsa cuando el verdadero valor del porcentaje que prefieren el producto es 0.10?

Rp. a) $\alpha=P[X<15/p=0.2]=P[Z<-1.25]=0.1056$, b) $((K/100)-0.2)/0.4=-1.645$, $K=13.42$, c) $\beta=P[X\geq C/p=0.1]=P[Z\geq 1.14]=0.1271$.

36. La SUMAT afirma que el 70% de los contribuyentes pagan sus impuestos correctamente. Si menos de 52 contribuyentes de una muestra de 80 contribuyentes pagan sus impuestos correctamente, se rechaza la afirmación. Se acepta en caso contrario.

a) Plantee las hipótesis de esta prueba

b) ¿Cuál es el nivel de significación de la prueba?

c) ¿En qué porcentaje la prueba puede detectar una diferencia de 20% por debajo de lo indicado en la hipótesis nula?.

Rp. a) H_0: p=0.70, H_1: p<0.70, b) α=0.1635, b) $\beta=P[\,\overline{P} \geq 0.65\,/\,p = 0.50]=P[Z \geq 2.68]=$ 0.0037. La prueba detecta la diferencia el $1-\beta$= 99.63% de las veces.

37. De una lista de 2,000 clientes de un hipermercado que pagan a plazos, se seleccionó una muestra aleatoria para obtener opinión acerca del servicio. En la muestra se halló que 215 no tienen quejas del servicio, 25 tienen quejas y 10 no opinan al respecto. Tradicionalmente el 5% tenían quejas del servicio, sin embargo se cree que ahora este porcentaje aumentó. ¿Cuál es la situación actual si se quiere una probabilidad de 0.008 de cometer error tipo I?.

Rp. Población finita H_0: p=0.05, H_1: p>0.05, $RC=\{Z>2.41\}$, z_k=3.88, se rechaza H_0.

38. Un empleado del banco "TRABAJA" ha revisado 2000 créditos. Luego, un auditor seleccionó al azar 400 de tales créditos y encontró que en 20 de ellas había errores. Considerando como satisfactoria hasta un 3% de créditos con error y en el nivel de significación de 3%, ¿puede admitirse como satisfactorio el trabajo del empleado?.

Rp. Población finita: $H_0 : p \leq 0.03$, $H_1 : p > 0.03$, $z_k = 2.62$, $R.C. =]1.88, +\infty[$, se rechaza H_0.

39. El administrador del banco "CREDITOS" afirma que el 10% de los clientes hacen operaciones diarias por más de $10,000. Se va a diseñar una prueba de hipótesis para el porcentaje. Halle el tamaño de la muestra y el valor crítico de la prueba si se desea que la probabilidad de cometer error tipo I sea 0.0228 y que el riesgo de tomar una decisión equivocada sea 0.0329 cuando la proporción de clientes que hacen operaciones por más de $10,000 sea realmente 5% .

Rp. $n \cong 400$, $K \cong 0.07$

40. El jefe de control de la firma de confecciones "INDIO" afirma que sólo el 1% de las prendas producidas por ellos no satisface el control de calidad. Se desea comprobar esta hipótesis contra la alternativa que supone es un 3% el porcentaje de prendas que no pasan el control de calidad. Si se quiere correr un riesgo de no más de 5 casos de 100 de rechazar la afirmación del jefe de control si es realmente verdadera y de correr un riesgo de no más de 1 caso de 100 de aceptar la afirmación del jefe de control cuando la alternativa es realmente cierta, ¿cuál debe ser el tamaño de la muestra a seleccionar y cuál la región crítica o de rechazo de la afirmación del jefe de planta?.

Rp. $n \cong 788$, R.C.= $]0.016, +\infty[$.

Varianzas

41. Los diámetros en centímetros de las piezas que produce un torno tiene distribución normal con desviación estándar de 0.25. En un reciente control una muestra aleatoria de 20 piezas dio una desviación estándar de 0.32 En el nivel de significación 0.05, ¿es el aparente incremento de variabilidad significativo?.

Rp. H_0: $\sigma^2 = (0.25)^2$, H_1: $\sigma^2 > (0.25)^2$, $X \sim \chi^2(19)$, $RC=\{X>30.14\}$, $\chi^2 = 19\times(0.32)^2 / (0.25)^2 = 31.13$, se rechaza H_0, si.

42. El peso en gramos de los envases no retornables de gaseosa tiene distribución normal con una media de 10 y una varianza igual a 0.25. Para comprobar el valor de la varianza se escogió una muestra aleatoria de 16 envases resultando una varianza de 0.20. En el nivel de significación del 5%, ¿es válido inferir que la varianza de los pesos de tales envases es menor que 0.25?

Rp. H_0: $\sigma^2=0.25$, H_1: $\sigma^2<0.25$, $X \sim \chi^2(15)$, $RC=\{X<7.26\}$, $x_k=12$, se acepta H_0.

43. Las ventas mensuales en soles de filetes de atún en todas los mercados, tienen distribución normal con una varianza de 400. Para realizar una prueba del valor de la varianza se escogió una muestra aleatoria de 13 tiendas y se encuentra que las ventas del mes dan una varianza de 360. Al nivel de significación del 5%, ¿hay razón suficiente para concluir que la varianza de la población es diferente de 400?.

Rp. H_0: $\sigma^2=400$, H_1: $\sigma^2\neq 400$, $X \sim \chi^2(12)$, $RC=\{X<4.40$ o $X>23.34\}$, $x_k=10.8$, se acepta H_0.

44. Los salarios en dólares del personal de las compañías A y B se distribuyen según el modelo de probabilidad normal con igual media. Para determinar cuál de ellas tiene salarios más homogéneos, se escogió una muestra aleatoria de 10 salarios de A, y 9 de B resultando las varianzas 100 y 225 respectivamente. En el nivel de significación $\alpha=0.01$, ¿hay razón suficiente para decidir que en la compañía A los salarios son más homogéneos?

Rp. H_0: $(\sigma_1)^2=(\sigma_2)^2$, H_1: $(\sigma_1)^2<(\sigma_2)^2$, $F \sim F(9,8)$, $RC=\{F<0.183\}$, $f_k=0.44$, se acepta H_0.

45. El jefe de logística de cerámicas "LOZA" tiene que escoger entre dos marcas A y B de máquinas para su planta de producción. El sabe que cada marca tiene un tiempo de producción por pieza cuya distribución es normal. Se le permitió probar ambas máquinas durante un periodo de prueba para luego escoger 10 tiempos al azar para cada un de ellas, resultando los siguientes tiempos en segundos:

Máquina A: 40, 49, 47, 42, 48, 38, 44, 49, 50, 37

Máquina B: 40, 41, 39, 40, 38, 42, 43, 37, 38, 41

a) En el nivel de significación de 0.05 y en una prueba bilateral, ¿se podría concluir que las varianzas poblacionales son iguales?. ¿Qué marca de máquina debería adquirir?
b) Determine la probabilidad P de significación unilateral

Rp. a) H_0: $(\sigma_1)^2=(\sigma_2)^2$, H_1: $(\sigma_1)^2\neq(\sigma_2)^2$, $F\sim F(9,9)$, $RC=\{F<0.248$ o $F>4.03\}$, $f_k=(4.88)^2/(1.912)^2=6.517$, se rechaza H_0, en una prueba unilateral al 5% se infiere que la máquina B tiene menor dispersión, b) Signif . $P=0.005$.

Diferencia de dos medias

46. Un estudio estadístico sobre el uso de cajeros automáticos indica que el monto diario (en dólares) de los movimientos tanto para hombres y mujeres tienen distribución normal con la misma media y con varianzas respectivas de 64 y 49. Sin embargo la inferencia respecto a la igualdad de las medias es poco creíble. Para investigar más al respecto, se seleccionaron aleatoriamente los montos de los movimientos de 20 hombres y 25 mujeres dando las medias respectivas de 200 y 205. Para el nivel de significación del 1%, ¿se puede concluir que las medias de las dos poblaciones de montos son diferentes?.

Rp. H_0:$\mu_1=\mu_2$, H_1:$\mu_1\neq\mu_2$ $z_k=-2.2$, $RC=\{Z<-2.575$ o $Z>2.575\}$ se acepta H_0. No.

47. Un inversionista está por decidir entre dos localidades para abrir un centro comercial. Para esto debe probar la hipótesis de que hay diferencia en la media de los ingresos mensuales de los hogares de las dos provincias. Se escogió una muestra aleatoria de cada lugar y se obtiene la tabla de resultados en dólares:

	Localidad	
	A	B
Tamaño muestral	300	400
Media muestral	400	420
Varianza muestral	8100	14400

a) ¿Qué estadística es la apropiada para esta prueba de hipótesis?
b) Para un nivel de significación de 0.05, ¿puede el inversionista concluir que le es indiferente construir en cualquiera de las dos localidades?, si no es así, ¿en cuál de las localidades debería abrir el centro comercial?

Rp. a) Z, b) $z_k=-2.52$, $R.C=\{Z<-1.96$ o $Z>1.96\}$, $\mu_1\neq\mu_2$, además $\mu_1<\mu_2$, en B.

48. Un analista financiero está interesado en comparar los niveles de rendimiento, en puntos porcentuales, de dos empresas de sectores diferentes1 y 2. El sabe que las tasas de rendimiento de cada una de estas empresas tiene distribución normal. Seleccionó al azar 16 acciones de cada una de las empresas y observó

las tasas de rendimiento. Las tasas de rendimiento dieron las medias 45 y 38, y las varianzas 128 y 64 respectivamente para las empresas 1 y 2.
Al nivel de significación 0.05

a) ¿Son diferentes las dos varianzas poblacionales de las tasas de rendimiento?
b) ¿Es la tasa de rendimiento promedio de la empresa 1 mayor que la de la empresa 2?.

Rp. a) H_0: $(\sigma_1)^2=(\sigma_2)^2$, H_1: $(\sigma_1)^2\neq(\sigma_2)^2$, $F\sim F(15,15)$, $RC=\{F<0.35$ o $F>2.86\}$, $f_k=2$, se acepta H_0
b) H_0:$\mu_1=\mu_2$, H_1:$\mu_1>\mu_2$, gl=30, error estándar=3.46, $t_k=2.023$, , gl=30, $RC=\{T>1.697\}$, Si

49. El gerente de compras de la empresa de transportes "CARGA" debe decidir por dos marcas A y B de bujías para su flota de camiones. El sabe que las vidas útiles en km., para cada marca de bujía tienen distribución normal. La vida útil de una muestra aleatoria de 10 bujías de la marca A, dio una media de 8000 y una varianza de 5600. La vida útil de una muestra aleatoria de 9 bujías de la marca B dio una media de 7900 y una varianza de 810.
Al nivel de significación 0.05

a) Realice una prueba bilateral de homogeneidad de las varianzas?
b) ¿Por cuál de las dos marcas de bujía debería decidir el gerente?.

Rp. a) H_0: $(\sigma_1)^2=(\sigma_2)^2$, H_1: $(\sigma_1)^2\neq(\sigma_2)^2$, $F\sim F(9,8)$, $RC=\{F<0.244$ o $F>4.36\}$, $f_k=6.91$, se rechaza H_0 b) H_0:$\mu_1=\mu_2$, H_1:$\mu_1>\mu_2$, $t_k=3.922$, gl=11.78≅12, $RC=\{T>1.782\}$, A.

50. El grupo "NATURA" lanza la publicidad de su producto de fibra natural afirmando que su consumo durante un mes da como resultado la pérdida de peso. Una muestra aleatoria de 12 personas que consumieron el producto reveló un peso medio de 62 kg antes, un peso medio de 58 kg después de un mes de iniciado el consumo y una desviación estándar de las diferencias de pesos $\hat{s}_d = 5$ kg..

Suponga que la diferencia de los pesos tiene distribución normal.
En el nivel de significación de 0.01.

a) ¿Se debería concluir que el producto es efectivo?
b) Halle el valor de la probabilidad P de la prueba.

Rp. H_0:$\mu_1-\mu_2=0$, H_1:$\mu_1<\mu_2$, $\bar{d}=4$, Error estándar=1.44, $t_K=2.778$, gl=11, , a) $RC=\{T>2.718\}$, se rechaza H_0. b) $P=P[T>2.778]=0.00898$

51. El agente de compras de la compañía "PC" se vio confrontado con dos marcas de computadoras para su adquisición. Se le permitió probar ambas marcas asignando una misma tarea a 50 máquinas de cada marca, resultando las medias respectivas 55 y 50 minutos. Suponga las dos poblaciones tienen varianza homogénea igual a 100. Para el nivel de significación $\alpha=0.05$:

a) ¿Excede el tiempo promedio de la marca 1 al de la marca 2 en al menos 9 minutos?.
b) Halle la potencia de la prueba cuando la diferencia real entre promedios de tiempo de marca 1 menos marca 2 sea 3 minutos.
c) ¿Qué tan grande debe ser la muestra si la potencia de la prueba es 0.95, cuando la diferencia real entre promedios de tiempo marca 1 menos marca 2 es 3 minutos.

Rp. a) $H_0:\mu_1-\mu_2\geq 9$, $H_1:\mu_1-\mu_2<9$, $ES=2$, $z_k=-2$, $RC=\{Z<-1.645\}$ se rechaza H_0., b) $1-\beta=0.9131$, c) $n=105$

Utilizando un paquete de computo (por ejemplo *MCEST*) resolver:

52. Dos profesores enseñan a dos secciones A y B el mismo curso de matemáticas. Para comparar los promedios en las calificaciones obtenidas con los dos profesores, se escogieron dos muestras aleatorias independientes de 9 notas de A y 8 notas de B dando los siguientes resultados:

Grupo A: 02, 18, 10, 20, 17, 05, 12, 16, 11.
Grupo B: 12, 16, 09, 16, 12, 13, 11, 10.

Suponga que las calificaciones con cada uno de los dos profesores se distribuyen normalmente
Con un nivel de significación de 0.05,

a) ¿Se podría concluir que son homogéneas las varianzas de las calificaciones con los dos profesores?
b) ¿Es la calificación promedio de A más alta que la de B?

Rp. a) $s_A=6.062$, $s_B=2.56$, $F=5.607$, gl: 8, 7, sigf.bilat=0.0348, se acepta H_1: $(\sigma_1)^2\neq(\sigma_2)^2$. medias: 12.33 y 12.37, $t_k=-0.0188$, g.l.=11.02≅11. sigf.unilateral=0.4925, aceptar $H_0:\mu_1=\mu_2$.

53. La empresa de transporte terrestre de pasajeros "BUS" está por decidir si compra la marca A o la marca B de llantas para su flota de ómnibus. Se sabe que el rendimiento de cada marca tiene distribución normal. Se probaron dos muestras independientes de 9 llantas de las marcas A y B resultando los siguientes rendimientos en kilómetros:

Marca A: 32000, 30000, 33000, 31000, 32000, 35000, 34000, 35000, 31000

Marca B: 35000, 37000, 36000, 38000, 37000, 39000, 32000, 33000, 40000

Con un nivel de significación de 0.01

a) ¿Es razonable concluir que son significativamente diferentes las varianzas de los rendimientos?

b) ¿Es posible concluir que las dos marcas rinden igual?.

Rp. a) F=2.136, sigf.bilateral=0.3037 se acepta $H_0:(\sigma_1)^2=(\sigma_2)^2$,
b) t_k=−3.54, g.l.=16, sigf.bil=0.0027, se rechaza $H_0:\mu_1=\mu_2$.

54. Las ventas medias semanales de las llantas PS214 en dos tiendas A y B de servicios, son aproximadamente iguales. Sin embargo el gerente de ventas de la tienda B cree que sus ventas son más consistentes. A continuación se presenta el número de llantas PS214 que se vendieron en las últimas 10 semanas en la tienda A y durante las últimas 11 semanas en la tienda B:

Tienda A: 32, 35, 34, 35, 32, 30, 33, 31, 31, 33
Tienda B: 39, 38, 40, 42, 45, 44, 35, 32, 36, 38, 37

Suponga que tales ventas en cada tienda tienen distribución normal. En el nivel de significación α=0.05
a) ¿Son homogéneas las varianzas de las ventas semanales?
b) ¿Está usted de acuerdo con el gerente de al tienda B?.

Rp. a) F=5.12, gl:10, 9, sigf.bil=0.022 se rechaza $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=−4.758, g.l=14.038≅14, sigf.unil=0.00015, se acepta $\mu_1<\mu_2$ B vende más.

55. En un estudio para comparar la venta diaria de arroz embolsado en los hipermercados 1 y 2 se escogieron el monto de las ventas de 9 y 8 días al azar respectivamente de 1 y 2 resultando los siguientes datos en soles.:

Mercado 1: 1500, 1700, 1600, 1800, 1700, 1900, 1200, 1300, 1400
Mercado 2: 1200, 1000, 1300, 1100, 1200, 1500, 1400, 1500

Suponga que tales ventas en cada uno de los supermercados tienen distribución normal.
Utilizando el nivel de significación del 1%,
a) ¿Se puede concluir que las varianzas de todas las ventas son iguales?
b) Utilice una prueba unilateral, para determinar si existe alguna diferencia significativa en las ventas medias?.

Rp. F=1.6382, gl:8,7, sigf.bil=0.52917 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=2.8295, g.l=15, ES=107.217, sigf.bil=2.8728, sigf.unil=0.006, se acepta $\mu_1>\mu_2$.

56. El gerente de ventas de pantalones "INCA" quiere saber si una reducción del 5% en el precio de su producto es suficiente para aumentar sus ventas. Para comprobar esta hipótesis el fabricante seleccionó en forma aleatoria 10 sucursales donde se vendió el producto a precio normal y otras 10 sucursales donde se vendió a precio de oferta. El número de unidades vendidas durante la semana pasada fue:

Precio de oferta: 55, 56, 57, 56, 58, 53, 54, 59, 60, 57
Precio normal: 50, 45, 49, 50, 38, 58, 63, 37, 48, 85

Suponga que cada una de tales ventas se distribuyen aproximadamente normal. En el nivel de significación de 0.05,

a) ¿Se puede concluir que son iguales las varianzas de los precios?
b) ¿Se puede inferir que la reducción del precio aumenta las ventas?.

Rp. F=0.0243, gl:9, 9, sigf.bil=0.0000 se acepta $(\sigma_1)^2 \neq (\sigma_2)^2$, b) t_k=0.9416, g.l=9.4375≅9, sigf.unil=0.18493, se acepta $\mu_1=\mu_2$.

57. Los salarios mensuales de los empleados de dos grandes empresas manufactureras A y B se distribuyen aproximadamente normal con medias iguales. Sin embargo los empleados de la empresa B creen tener los mejores salarios. Dos muestras aleatorias independientes de 8 empleados de A y de 9 empleados de B dieron los siguientes salarios en nuevos soles:

Muestra A: 3400, 3500, 3100, 3200, 3000, 3300, 3100, 3200
Muestra B: 3800, 3700, 3900, 3500, 3700, 3600, 3200, 3300, 4000

Al nivel de significación del 5%,

a) ¿Se puede concluir que las varianzas de los salarios son iguales?
b) ¿Es razonable concluir que los empleados de la empresa B están mejor pagados?. Realice una prueba unilateral.

Rp. a) F=0.39795, gl:7, 8, sigf.bil=0.242 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=−3.746, g.l=15, ES=109.015, sigf.unilateral=0.0009, se acepta $\mu_1<\mu_2$.

58. Se realiza el control en los frascos de 300 gramos de un producto que tiene solo dos componentes A y B en iguales cantidades promedio. Se sabe que cada componente del producto tiene distribución normal. Una muestra aleatoria de 10 frascos ha dado los siguientes porcentajes de la componente A

48%, 52%, 49%, 55%, 62%, 51%, 53%, 54%, 55%, 56%

Para un nivel de significación de 0.05.

a) Determine si las varianzas de los contenidos de las dos componentes son homogéneas.
b) ¿Son diferentes los promedios de los contenidos de las dos componentes?.

Rp. a) A: 144, 156, 147, 165, 186, 153, 159, 162, 165, 168, B: 156, 144, 153, 135, 114, 147, 141, 138, 135, 132, F=1, gl:9,9, sigf.bil=1.00 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=3.934, g.l=18, ES=5.34, sigf.bil=0.001, se rechaza $\mu_1=\mu_2$.

59. La compañía agroindustrial "MERMELADA" envasa uno de sus productos en frascos de 500 gramos. En la etiqueta de cada frasco se afirma que su contenido es en promedio 70% de fresa y 30% de piña. Se sabe que el contenido de cada una de las dos componentes tiene distribución normal. Un investigador examinó el contenido de 9 frascos escogidos al azar, resultando los siguientes **porcentajes** de fresa:

71, 67, 65, 63, 58, 72, 68, 64, 65

Para un nivel de significación de 0.01.

a) ¿Son iguales las varianzas de los contenidos de las dos componentes?.

b) ¿Dan los datos prueba suficiente de que la diferencia entre el contenido promedio de fresa y de piña de los frascos sea mayor a 100 gramos?

Rp. a) FR: 355, 335, 325, 315, 290, 360, 340, 320, 325, PI: 145, 165, 175, 185, 210, 140, 160, 180, 175, F=1, gl:8,8, sigf.bil=1.00 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=5.87, g.l=16, ES=10.031, sigf.unil=0.000, se rechaza $\mu_1=\mu_2+100$.

60. Un alumno de estadística aplicada quiere comprar los precios que calculan los tasadores 1 y 2 para automóviles usados. Seleccionó una muestra de 10 autos y pidió que ambos tasadores los valuaran. Los siguientes son los precios en cientos de dólares.

Automóvil	1	2	3	4	5	6	7	8	9	10
Tasador 1	54	51	28	48	35	26	54	48	56	48
Tasador 2	49	47	30	43	27	31	52	41	57	38

Suponga que los precios de cada tasador tiene distribución normal.
Para el nivel de significación de 0.01.¿Se puede afirmar que el precio promedio del tasador 1 es mayor que el promedio de precios del tasador 2?

Rp. $\bar{d}=3.3$, $s_d=4.768$, ES=1.506, gl=9, t_K=2.19, $RA=\{T\leq 2.821\}$, Sigf.unil=0.028, no se rechaza $\mu_1=\mu_2$.

61. La compañía de transporte interprovincial "CARGA" debe decidir si compra la marca A o la marca B de neumáticos para su flota de camiones. Para estimar la diferencia entre las dos marcas asignó un neumático de cada marca a las ruedas delanteras de 12 ómnibus y se registraron en miles de kilómetros las siguientes distancias:

Camión	1	2	3	4	5	6	7	8	9	10	11	12
Marca A	50	47	38	44	35	36	44	48	46	48	49	51
Marca B	45	43	30	39	35	31	42	44	37	46	48	52

Utilizando un nivel de significación del 5%, ¿se puede concluir que los promedios de rendimiento son iguales en ambas marcas con una prueba bilateral?. Suponga que las diferencias de las distancias se distribuyen en forma normal.

Rp. $\bar{d}=3.67$, $s_d=3.025$, ES=0.873, gl=11, t_K=4.199, $RA=\{-2.201\leq T\leq 2.201\}$ Sigf.bil=0.001, se rechaza $\mu_1=\mu_2$.

62. La gerencia de ventas de una cadena de mercados diseñó un plan de incentivos para sus vendedores. A fin de evaluar este plan se seleccionaron doce vendedores al azar y se registraron sus ventas diarias en dólares antes y después de aplicar el plan:

Vendedor	1	2	3	4	5	6	7	8	9	10	11	12
Antes	89	87	70	83	67	71	92	81	97	78	94	79
Después	94	91	68	88	75	66	94	88	96	88	95	87

Suponga que las ventas antes y después del plan tienen distribución normal. Utilice el nivel de significación 0.05, para determinar si hubo algún incremento significativo en las ventas de todos sus vendedores debido a la aplicación del plan.

Rp. H_0:$\mu_1=\mu_2$, H_1:$\mu_1<\mu_2$, $d=x_1-x_2$, $\overline{d}=-3.5$, $s_d=4.5825$, $t_K=-2.6457$, $RC=\{T<-1.796\}$, Significación unilatral=0.011379, se rechaza H_0.

Diferencia de dos proporciones

63. Una empresa de estudios de mercado quiere saber si un producto promocionado a nivel nacional lo adquieren los hombres en mayor porcentaje que las mujeres. Para esto se escogieron dos muestras aleatorias independientes de 900 hombres y 800 mujeres resultando que 270 hombres y 200 mujeres adquirieron el producto.

a) Plantee las hipótesis nula y alternativa

b) Determine la proporción conjunta

c) Al nivel de significación del 5%, ¿cuál es su decisión respecto a la hipótesis nula?.

d) Determine el valor de la probabilidad P.

Rp. a) H_0:$p_1=p_2$, H_1:$p_1>p_2$, b) 0.27647, c) z_k=2.3, $RC=\{Z>1.645\}$ se rechaza H_0, d) 0.0107

64. En una encuesta de opinión realizada en Lima, en el mes 1 una muestra aleatoria de 200 ciudadanos de Lima indicó que 20 de ellos se abstienen de opinar. En el mes 2 otra muestra aleatoria de 200 ciudadanos demostró que 12 de ellos se abstienen de opinar sobre el mismo asunto. En el nivel de significación 0.05, verificar la afirmación de que la diferencia p_1-p_2 es menor que 5% donde p_1 y p_2 son las proporciones de todos los todos los ciudadanos que se abstienen de opinar.

Rp. H_0:$p_1-p_2=0.05$, H_1:$p_1-p_2<0.05$, $z_k=-0.37$, $RC=\{Z<-1.645\}$, no se rechaza H_0.

65. En una muestra de 500 hogares de Trujillo se encontró que 50 de ellos estaban viendo vía satélite un programa especial de televisión. En otra muestra de 400 hogares de Tarapoto se encontró que 28 de ellos estaban viendo el mismo programa especial. En el nivel de significación 0.05, ¿puede rechazarse la suposición del patrocinador de que el porcentaje de hogares que están observando el programa especial no es el mismo en las dos ciudades?.

Rp. $H_0{:}p_1=p_2$, $H_1{:}p_1\neq p_2$, $z_k=1.59$, $RA=\{-1.96\leq Z\leq 1.96\}$ se acepta H_0.

66. En un estudio de mercado para determinar el rating de los programas de TV del mediodía una muestra aleatoria de 400 hogares de cierta comunidad reveló que 80 están sintonizando el programa B de TV, 120 sintonizan el programa G y el resto sintonizan otra cosa. ¿Es la proporción global de televidentes que sintonizan el programa B igual al que sintonizan G?. Utilice $\alpha= 0.01$ y una prueba bilateral.

Rp. $H_0{:}p_1=p_2$, $H_1{:}p_1\neq p_2$, $z_k=-3.27$, $RA=\{-2.575\leq Z\leq 2.575\}$ se rechaza H_0.

67. La agencia de publicidad "RDT" realizó un estudio para comparar la efectividad de un anuncio en la radio en dos distritos. Después de difundir dicho aviso, se realizó una encuesta telefónica con 600 personas seleccionadas al azar, que viven en cada uno de los dos distritos resultando las proporciones: 20% y 18% respectivamente para el primero y el segundo distrito. En el nivel de significación del 5%, ¿es posible concluir que la proporción de todas las personas que escucharon dicho aviso en el primer distrito es superior a la del segundo distrito?.

Rp. $H_0{:}p_1=p_2$, $H_1{:}p_1>p_2$, $z_k=0.88$, $RC=\{Z>1.645\}$ se acepta H_0.

68. Para probar la eficacia de dos nuevos insecticidas en la protección contra plagas de las viñas de San Antonio en San Martín, se seleccionaron al azar 80 plantas de uvas para rociarlo con el insecticida A y 50 plantas de uvas para rociarlo con el insecticida B. Cuando maduraron las uvas se encontró que 6 y 5 plantas de uvas rociadas con A y B respectivamente tenían plagas. Con un nivel de significación del 5%, ¿se puede concluir que el insecticida A es más eficaz?

Rp. $H_0{:}p_1=p_2$, $H_1{:}p_1<p_2$, $z_k=-0.5$, $RC=\{Z<-1.645\}$ se acepta H_0.

69. El gerente de compras de una compañía está evaluando dos marcas de equipo para fabricar un artículo. Examinó una muestra aleatoria de tamaño 50 para la primera marca y encontró que 5 de ellos tenían defectos. Controló otra muestra aleatoria de tamaño 80 para la segunda marca y encontró que 6 de ellos tenían defectos. Los manuales de los equipos indican que el porcentaje de fabricación defectuosa del total es la misma para las dos marcas de equipo. Sin embargo, como la primera cuesta bastante menos, el gerente de compras le otorga a esa marca el beneficio de la duda y afirma que la primera tiene mayor porcentaje de

producción defectuosa. En el nivel de significación de 0.05, ¿cuál es la conclusión de usted?.

Rp. H_0: $p_1 \leq p_2$, , H_1: $p_1 > p_2$, $\hat{p} = 0.085$, ES=0.051, z_K=0.49, $R.C.$=]1.645, +∞[, se acepta H_0.

70 *. Use un paquete de computo estadístico para resolver el siguiente problema:

Con referencia a la hoja de cálculo del *estudio socioeconómico de universitarios de Lima* (ver apéndice)

a) En el nivel de significación de 0.05, ¿es posible concluir que el ingreso medio mensual de las familias de los universitarios de Lima es superior a 3300?. ¿Cuál es el valor de la probabilidad P?.

b) En el nivel de significación de 0.01, ¿es posible concluir que menos del 60% de los universitarios provienen de colegios nacionales?. ¿Cuál es el valor de P?

Rp. a) Z=1.328, No, p=0.0921, b) Z=−1.706, No, p=0.044

Capítulo 11

LA PRUEBA DE CHI-CUADRADO

Introducción

Los métodos de pruebas de hipótesis expuestas en el capitulo anterior se han basado en el supuesto que las muestras se han escogido de poblaciones que tienen distribuciones conocidas, como, por ejemplo la distribución normal.

Cada distribución depende de uno o más parámetros, por esta razón estas técnicas estadísticas se denominan **métodos** ***paramétricos***.

Las pruebas que consisten en sacar conclusiones directamente de las observaciones muestrales, sin formular los supuestos acerca del tipo de distribución de la población de la que proviene, se denominan **pruebas** ***no paramétricas*** o *de libre distribución*.

En este capitulo se cubre los métodos no paramétricos que incluyen la aplicación de la distribución chi-cuadrado en ***pruebas de bondad de ajuste.***
Y en pruebas de hipótesis que se relacionan con *tablas de contingencia* tales como:

* *prueba de independencia de dos variables estadísticas*,
* *prueba de homogeneidad de muestras*, y
* *prueba de igualdad de dos o más proporciones poblacionales.*

Se recomienda al lector que los ejemplos y ejercicios sean resueltos utilizando un paquete de computo, por ejemplo, el paquete estadístico didáctico *MCEST*, creado por el autor de este libro.

En el capitulo 15 se dan los métodos no paramétricos de promedios.

11.1 Pruebas de bondad de ajuste

La prueba de bondad de ajuste consiste en determinar si una población tiene una distribución teórica o hipotética específica a partir de los resultados de una muestra aleatoria escogida de esa población.

La hipótesis nula en una prueba de bondad de ajuste consiste en afirmar que la distribución de frecuencias observadas concuerda con el modelo de probabilidad esperado de las frecuencias en un conjunto de clases o categorías.

La distribución de probabilidad teórica o esperada puede referirse a la distribución uniforme, a la binomial, a la Poisson, o a la normal, etc.

Supongamos que un experimento produce k eventos mutuamente excluyentes: $E_1, E_2, ..., E_k$ (llamados también clases o categorías) y que la probabilidad de que ocurra el evento E_i sea p_i, donde:

$$\text{para cada } i \text{ es: }, \; p_i > 0, \text{ y } \sum_{i=1}^{k} p_i = 1$$

Supongamos también que se repite n veces el experimento aleatorio y sean X_i el número de veces que ocurre el evento E_i. Cada X_i es una variable aleatoria cuyo valor x_i es la *frecuencia observada* en la i-ésima clase. El conjunto de valores observados constituye la muestra aleatoria de tamaño n cuyos resultados se muestran en la siguiente tabla:

Categorías	$E_1 \quad E_2 \quad \ldots \quad E_k$	Total
frecuencias observadas	$x_1 \quad x_2 \quad \ldots \quad x_k$	n

En 1900, Karl Pearson ha demostrado que cuando:

$$Z_i = \frac{X_i - np_i}{\sqrt{np_i}}, \quad i = 1, 2, ..., k$$

y cuando n tiende al infinito, la variable aleatoria:

$$W = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

se aproxima a una distribución chi-cuadrado con $k-1$ grados de libertad.

Es conveniente simbolizar esta aproximación por:

$$W = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

en donde, O_i es la *frecuencia observada* X_i y e_i es la *frecuencia esperada* np_i.

Se puede pues, utilizar la estadística W como una medida de la discrepancia entre las frecuencias observadas y esperadas. Si $W = 0$, las frecuencias observadas y esperadas concuerdan exactamente, mientras que si $W > 0$, no coinciden exactamente. A mayores valores de W, mayores son las discrepancias entre valores observados y esperados.

La prueba

Las **hipótesis** nula y alternativa de la prueba de bondad de ajuste son respectivamente:

H_0 : La distribución de frecuencias de la muestra concuerda con la distribución teórica (o hipotética) propuesta

H_1 : La distribución de la muestra no concuerda con la distribución teórica.

Las **frecuencias esperadas** e_i se calculan a partir de la distribución teórica que se propone en la hipótesis nula H_0. Para que exista una buena aproximación a la distribución chi-cuadrado debemos establecer que todas las frecuencias esperadas deben ser no menos de 5. Cuando alguna clase tiene frecuencia esperada menor que 5 se agrupan dos o más clases adyacentes en una sola de manera que la frecuencia esperada sea mayor o igual que 5. Esto implica la reducción de los grados de libertad.

Con las frecuencias observadas o_i y las frecuencias esperadas e_i, se calcula la **estadística:**

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

en donde k es el número de clases que resultan con frecuencias esperadas mayores o iguales que 5.

Dado el **nivel de significación** α y para $k-1$ grados de libertad, en la tabla chi-cuadrado se halla el número $c = \chi^2_{k-1,1-\alpha}$ que viene a ser el valor crítico de la prueba.

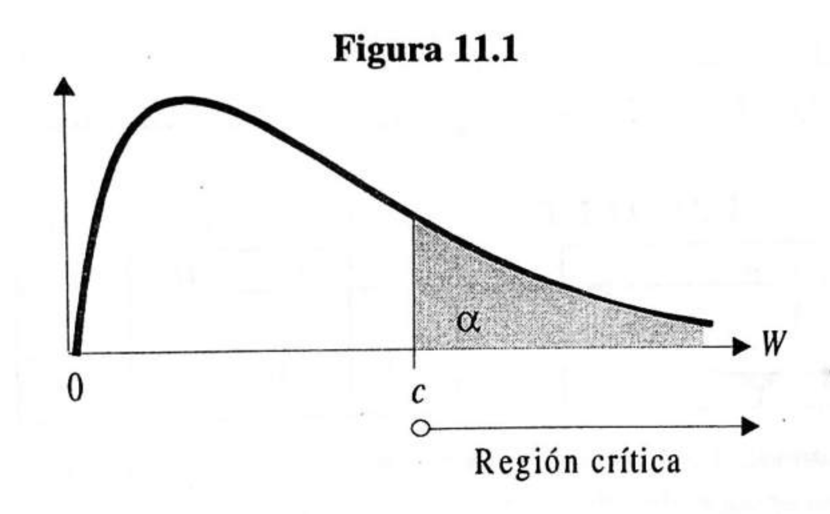
La **región crítica** de la prueba de la hipótesis nula H_0 en la distribución de $W \sim \chi^2_{k-1}$ (ver figura 11.1) es el intervalo abierto

$$R.\,C. =]\, c, +\infty\, [$$

La **regla de decisión** para esta prueba es: Rechazar H_0 si $\chi^2_{cal} > c$, en caso contrario, se aceptará o al menos no se rechazará H_0.

En el modo *P*, del paquete *MCEST*, si $P = P[W > \chi^2_{cal}]$, se rechazará la hipótesis nula H_0 si $P < \alpha$. En caso contrario no se debe rechazar H_0.

Figura 11.1



NOTA 1. Si para calcular las frecuencias esperadas e_i se deben estimar m parámetros de la distribución teórica a partir de la distribución muestral, entonces, el número de grados de libertad es: $k-1-m$

NOTA 2. $\displaystyle\sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(o_i)^2}{e_i} - n$

EJEMPLO 11.1. (Ajuste a una distribución uniforme).

El gerente de ventas de un supermercado obtiene los siguientes datos de preferencias de cada una de las marcas detergentes: A, B, C, D, E, F :

Detergente	A	B	C	D	E	F
Número de amas de casa	28	27	32	34	30	29

Con estos resultados, ¿se puede concluir con un nivel de 0.05, que hay diferencias en las preferencias de las marcas de detergentes en la población de consumidores? .

SOLUCION.

Primero determinemos las frecuencias esperadas e_i a partir de la hipótesis nula. La hipótesis nula consiste en suponer que las preferencias de todos los consumidores por las marcas de detergentes tiene distribución teórica uniforme de 6 categorías

Si la hipótesis nula es verdadera, la probabilidad en cada categoría es :

$$p_i = \frac{1}{6}, \; i = 1, 2, 3, 4, 5, 6.$$

En consecuencia, la frecuencia esperada en cada categoría es:

$$e_i = np_i = 180(\frac{1}{6}) = 30 .$$

Las frecuencias observadas o_i y esperadas e_i se dan en la tabla 11.1:

Tabla 11.1. *Frecuencias observadas y esperadas*

Detergente	A	B	C	D	E	F	Total
Frecuencias observadas o_i	28	27	32	34	30	29	180
Frecuencias esperadas e_i	30	30	30	30	30	30	180

Note que cada una de las 6 frecuencias esperadas ($k = 6$) son no menores que 5. Por otro lado, no se ha estimado parámetro alguno ($m = 0$). Por lo tanto, el número de grados de libertad resulta es $k - 1 - m = 6 - 1 - 0 = 5$.

Continuando con la prueba se tiene:

1. *Hipótesis*:
   H_0 : No hay diferencias entre valores observados y esperados (o La distribución de la muestra se ajusta a la distribución uniforme de 6 celdas o categorías).
   H_1 : Si hay diferencias entre valores observados y esperados.
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística*: $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$, que se distribuye aproximadamente como chi-cuadrado con grados de libertad.
4. *Región crítica*. Para el nivel de significación 0.05 y 5 grados de libertad el valor crítico de la prueba es: $\chi^2_{0.95,5} = 11.07$.

Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor que 11.07. No se rechazará en caso contrario.

5. *Cálculos*. De la tabla 11.1 se obtiene:

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(28-30)^2}{30} + \frac{(27-30)^2}{30} + \frac{(32-30)^2}{30} + \frac{(34-30)^2}{30}$$

$$+ \frac{(30-30)^2}{30} + \frac{(29-30)^2}{30} = 1.133$$

6. *Decisión*. Dado que $\chi^2_{cal} = 1.133 < 11.07$, no se debería rechazar, H_0.

NOTA. Utilizando el paquete estadístico *MCEST* se obtiene la probabilidad $P=P[\chi^2(6)>1.133]=0.951$, probabilidad conocida como la **significación** de la prueba. En consecuencia, dado que $P > 0.05$, la decisión debería ser NO rechazar la hipótesis nula, H_0.

EJEMPLO 11.2. (Ajuste a una distribución binomial).

Durante 100 días un auditor revisa 5 declaraciones juradas cada día, seleccionándolas al azar. El número de declaraciones juradas erradas que encuentra por día tabula en una distribución de frecuencias, resultando:

No. de declaraciones erradas	0	1	2	3	4	5
Número de días	12	25	30	20	10	3

En el nivel de significación 0.05, ¿podemos concluir que esta distribución de frecuencias observadas concuerda con la distribución binomial $B(5{,}0.4)$?.

SOLUCION.

Sea X la variable aleatoria que denota el número de declaraciones juradas erradas que encuentra por día. La distribución teórica de X es la binomial $B(5, p)$.

El parámetro p se estima a partir de la distribución de frecuencias observadas, sabiendo que la media de la distribución teórica es $\mu = 5p$. (o $\hat{p} = \hat{\mu}/5$)

La estimación de μ es la media $\hat{\mu} = \bar{x}$ de la distribución de los datos:

$$\bar{x} = \frac{0\times12 + 1\times25 + 2\times30 + 3\times20 + 4\times10 + 5\times3}{12+25+30+20+10+3} = \frac{200}{100} = 2$$

Luego $\hat{p} = \hat{\mu}/5 = 2/5 = 0.4$.

Para calcular las frecuencias esperadas, suponemos que es verdadera la hipótesis nula: $X\sim B(5,\ 0.4)$.

Las frecuencias esperadas, entonces, son: $e_i = np_i = 100p_i$, donde,

$$p_i = P[X = i] = C_i^5 (0.4)^i (0.6)^{5-i}, \quad i = 0, 1, 2, 3, 4, 5$$

Las frecuencias observadas, las probabilidades binomiales y las frecuencias esperadas se dan en la tabla 11.2:

Tabla 11.2. *Frecuencias observadas y esperadas suponiendo binomial*

Número de caras	o_i	p_i	e_i
0	12	0.0778	7.78
1	25	0.2592	25.92
2	30	0.3456	34.56
3	20	0.2304	23.04
4	10 } 13	0.0768	7.68 } 8.70
5	3	0.01024	1.02
Total	100	1.00000	

Observe que hay una frecuencia esperada menor que 5. Se han juntado en una sola categoría las dos últimas. Resultando $k = 5$ clases con frecuencias esperadas mayores que 5. Además, se ha estimado un parámetro ($m = 1$).
Por lo tanto, el número de grados de libertad resultante es: $k - 1 - m = 5-1-1 = 3$.

Continuando con el proceso de la prueba de bondad de ajuste se tiene:

1. *Hipótesis*:
   H_0: No hay diferencia entre valores observados y esperados. (o la distribución de los datos observados concuerda con la binomial $B(5, 0.4)$.
   H_1: La distribución de los datos no concuerda con la binomial $B(5, 0.4)$.
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística*: $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$, cuya distribución es aproximadamente una chi-cuadrado con 3 grados de libertad:.
4. *Región crítica*. $\alpha = 0.05$ y 3 grados en la distribución de $\chi^2(3)$ se halla el valor crítico $\chi^2_{0.95,3} = 7.82$. Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor de 7.82. No se rechazará en caso contrario.
5. *Cálculos*. De la tabla 11.2 resulta:

$$\chi^2_{cal} = \frac{(12-7.78)^2}{7.8} + \frac{(25-25.92)^2}{25.9} + \cdots + \frac{(20-23.04)^2}{23.04} + \frac{(13-8.70)^2}{8.7} = 5.45$$

6. *Decisión*. Dado que $\chi^2_{cal} = 5.45 < 7.82$, no se debe rechazar la hipótesis nula.

NOTA. La probabilidad P de la prueba es: $P=P[\chi^2(3)>5.45]=0.14165$

EJEMPLO 11.3. (Ajuste a una distribución de Poisson).

En una fabrica de confecciones el número de defectos por unidad en una muestra de 150 camisas, ha dado la siguiente distribución de frecuencias:

Número de defectos	0	1	2	3	4	5	6
Número de objetos	58	55	22	10	4	1	0

En el nivel de significación de 0.05, ¿podemos concluir que la distribución de la muestra se ajusta a una distribución de Poisson?.

SOLUCION.

Sea X, el número de defectos por camisa. La hipótesis nula H_0: consiste en afirmar que la distribución teórica de X es Poisson $P(\lambda)$, donde el parámetro λ se debe estimar utilizando la distribución de los datos.

La estimación puntual de λ es la media $\hat{\lambda} = \bar{x}$, donde,

$$\bar{x} = \frac{0\times58+1\times55+2\times22+3\times10+4\times4+5\times1+6\times0}{58+55+22+10+4+1+0} = \frac{150}{150} = 1$$

Luego, las frecuencias esperadas son: $e_i = np_i = 150p_i$, donde,

$$p_i = P[X=i] = e^{-1}\frac{1^i}{i!}, \quad i = 0,1,2,3,4,5,6$$

Las frecuencias observadas, las probabilidades y las frecuencias esperadas se indican en la tabla 11.3.

Tabla 11.3. *Frecuencias observadas y esperadas suponiendo Poisson*

Número de caras	o_i	p_i	E_i
0	58	0.3679	55.18
1	55	0.3679	55.18
2	22	0.1839	27.59
3	10 } 15	0.0613	9.20 } 12.04
4	4	0.0153	2.30
5	1	0.0031	0.46
6	0	0.0005	0.08
Total	150	1.0000	

Observe que hay 3 frecuencias esperadas menores que 5. Se agrupan las 4 últimas categorías en una sola, resultando la frecuencia esperada 12.04, y la

frecuencia observada 15. De las 7 clases originales resultan $k = 4$ clases con frecuencias esperadas mayores 5. Además se ha estimado un parámetro ($m = 1$).

Luego, el número de grados de libertad es: $k - 1 - m = 4 - 1 - 1 = 2$

Continuando el proceso de la prueba de bondad de ajuste se tiene:

1. *Hipótesis*:
   H_0 : La distribución de los datos se ajusta a la Poisson $P(\lambda=1)$
   H_1 : La distribución de los datos no se ajusta a la Poisson $P(\lambda=1)$
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística*: $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$, cuya distribución es aproximadamente chi-cuadrado con 2 grados de libertad.
4. *Región crítica*. Para el nivel de significación $\alpha = 0.05$ y 2 grados de libertad en la tabla chi-cuadrado se encuentra. el valor crítico $\chi^2_{0.95,2} = 5.99$. Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor de 5.99.
5. *Cálculos*. De la tabla 11.3 resulta:

$$\chi^2_{cal} = \frac{(58 - 55.18)^2}{55.18} + \frac{(55 - 55.18)^2}{55.18} + \cdots + \frac{(15 - 12.04)^2}{12.04} = 2.01$$

6. *Decisión*. Dado que $2.01 < 5.99$, no se debe rechazar la hipótesis nula.

NOTA. La probabilidad P de la prueba es: $P=P[\chi^2(2)>2.01]=0.36604$

EJEMPLO 11.4. (Ajuste a una distribución normal).

En una prueba de aptitud aplicada a 50 alumnos, se han obtenido las siguientes calificaciones;

15	18	21	22	22	23	24	25	26	27
27	28	29	30	31	32	32	32	33	33
33	33	33	34	34	34	34	34	35	35
35	36	36	36	36	36	37	38	39	40
41	42	42	43	43	44	45	46	47	47

a) Construya una **distribución de frecuencias por intervalos** y un **diagrama de troncos y hojas** de estos datos.

b) Pruebe la bondad del ajuste entre las frecuencias observadas en los intervalos y las correspondientes frecuencias esperadas de una distribución normal, utilizando un nivel de significación de 0.05.

SOLUCION.

a) **Diagrama de troncos y hojas:**

Utilizando los primeros dígitos como troncos y los segundos dígitos como hojas, se obtiene un diagrama de sólo 4 trocos y, en consecuencia, no proporciona una imagen adecuada de la distribución. Para evitar este problema se debe aumentar el número de troncos en la tabla. Esto se consigue escribiendo el tronco dos veces, el primer tronco con hojas: 0, 1, 2, 3, 4 y el segundo tronco* con hojas 5, 6, 7, 8, 9. (ver referencia 16 de la bibliografía página 59)

Distribución de frecuencias por intervalos.

El número de intervalos lo da el número de Sturges: $K = 1 + 3.3\log(n) = 6.6 \cong 7$.

La amplitud de los intervalos es $A = (47 - 15)/7 = 4.57 \cong 5$

El **diagrama de doble tronco** y hojas así como la distribución por intervalos se dan en la tabla que sigue

Troncos	Hojas	Intervalos	Frecuencias
1	58	[15, 20[	2
2	12234	[20, 25[	5
2*	567789	[25, 30[	6
3	012223333344444	[30, 35[	15
3*	55566666789	[35, 40[	11
4	0122334	[40, 45[	7
4*	5677	[45, 50]	4

b) Sea X la variable que denota las calificaciones en la prueba de aptitud. Como queremos verificar si la distribución de los datos es normal con media μ y varianza σ^2, entonces debemos primero estimar los parámetros μ y σ^2. La estimación puntual de μ es $\bar{x} = 34$ y la estimación puntual de la desviación estándar σ es el número $\hat{s} = 7.64$.

Las frecuencias esperadas en cada intervalo: I_i, $i = 1, 2, 3, ..., k$, son:

$$e_i = np_i = 50p_i,$$

siendo la probabilidad,

$$p_i = P[L_i \leq X < U_i] = P\left[\frac{L_i - \mu}{\sigma} \leq Z < \frac{U_i - \mu}{\sigma}\right]$$

en donde, $Z = \dfrac{X - \mu}{\sigma}$ es $N(0,1)$, L_i y U_i son los límites inferior y superior respectivamente del intervalo.

Por ejemplo en la cuarta clase $(i = 4)$, se tiene la probabilidad

$$p_4 = P[30 \leq X < 35] = P\left[\frac{30-34}{7.64} \leq Z < \frac{35-34}{7.64}\right] = P[-0.52 \leq Z < 0.13] = 0.2502.$$

Luego la frecuencia esperada es: $e_4 = np_4 = 50 \times 0.2502 = 12.5$.

Las demás probabilidades y las frecuencias esperadas se obtienen de una manera similar. Estos resultados se indican en la tabla 11.4, donde se observa que hay 2 frecuencias esperadas menores que 5. Se ha combinado en una sola categoría las 2 primeras y las 2 últimas de las 7 clases originales resultando $k = 5$ clases con frecuencias esperadas mayores que 5. Además se han estimado dos parámetros μ y σ, ($m = 2$).

El número de grados de libertad es: $k - 1 - m = 5 - 1 - 2 = 2$

Tabla 11.4. *Frecuencias observadas y esperadas suponiendo normalidad*

Calificaciones	o_i	p_i	e_i
[15, 20[	2 } 7	0.0272	1.4 } 5.6
[20, 25[	5	0.0844	4.2
[25, 30[	6	0.1845	9.2
[30, 35[	15	0.2502	12.5
[35, 40[	11	0.2325	11.7
[40, 45[	7 } 11	0.1197	6.0 } 9.5
[45, 50]	4	0.0701	3.5
Total	50		

Siguiendo con el proceso de la prueba de bondad de ajuste se tiene:

1. *Hipótesis*:

   H_0: La distribución de los datos se ajusta a una distribución normal con media 34 y desviación estándar 7.64.

   H_1: La distribución de los datos no se ajusta a esa distribución normal
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística*: $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$, que se distribuye aproximadamente como chi-cuadrado con 2 grados de libertad.
4. *Región crítica*. Para el nivel de significación $\alpha = 0.05$ y 2 grados de libertad en la tabla chi-cuadrado se encuentra. el valor crítico $\chi^2_{0.95,2} = 5.99$. Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor de 5.99.
5. *Cálculos*. De la tabla 11.4 resulta:

$$\chi^2_{cal} = \frac{(7-5.6)^2}{5.6} + \frac{(6-9.2)^2}{9.2} + \frac{(15-12.5)^2}{12.5} + \frac{(11-11.7)^2}{11.7} + \frac{(11-9.5)^2}{9.5} = 2.24$$

6. *Decisión*. Dado que $2.24 < 5.99$, se debe aceptar que la distribución de los datos se ajusta a la normal N(34, $(7.64)^2$).

NOTA. La probabilidad P de la prueba es: $P=P[\chi^2(2)>2.24]=0.32627$.

NOTA. (Método gráfico y método de Kolmogorov-Smirnov del ajuste a la normal)

Existen otros métodos para determinar si un conjunto de datos provienen o no de una distribución normal. Entre ellos se tienen:

a) Los métodos gráficos Q-Q (cuartil-cuartil) normal y P-P (percentil-pecentil) normal

b) El método se prueba no paramétrico de Kolmogorov-Smirnov para una muestra.

Ilustraremos con un ejemplo el método Gráfico P-P normal y el de Kolmogorov-Smirnov que no requiere organizar los datos por intervalos.

En general, la prueba de Kolmogorov-Smirnov se utiliza para probar que una muestra de datos se ajusta a una distribución particular teórica (Uniforme, Binomial, Poison, geométrica, normal, exponencial etc.).

EJEMPLO 11.5

Determinar si es normal la población de la que ha sido extraída la muestra aleatoria simple:

17, 15, 8, 13, 9, 12, 10, 14, 11, 16

Utilice el nivel de significación $\alpha = 0.01$.

SOLUCION.

El procedimiento es el siguiente

1. Se ordenan los n datos y se obtiene la distribución de frecuencia acumulativa relativa $S_n(x_i)$. Las proporciones acumuladas están dadas por $S_n(x_i) = k/n$, donde k es el número de observaciones menor o igual que x_i.

2. Se obtienen las proporciones o probabilidades teóricas $F_i = P[X = x_i]$ de la distribución normal a partir de los datos tipificados: $z_i = (x_i - \overline{x})/\hat{s}$. Las probabilidades acumuladas observadas y esperadas se muestran en la tabla 11.5

Tabla 11.5. Probabilidades acumuladas observadas y esperadas suponiendo normalidad

Datos Ordenados x_i	Valores Tipificados z_i	Prob acum Observada $S_n(x_i)$	Prob acum Esperada F_i	Diferencias $\mid S_n(x_i) - F_i \mid$
8	−1.49	0.1	0.0681	0.0319
9	−1.16	0.2	0.1230	0.0770
10	−0.83	0.3	0.2033	0.0967
11	−0.50	0.4	0.3085	0.0915
12	−0.17	0.5	0.4325	0.0675
13	0.17	0.6	0.5675	0.0325
14	0.50	0.7	0.6915	0.0085
15	0.83	0.8	0.7967	0.0033
16	1.16	0.9	0.8770	0.0230
17	1.49	1.0	0.9319	0.0687

Media $\overline{x} = 12.5$, *Desviación estándar* $\hat{s} = 3.02765$.

3. Si los datos provienen de una distribución normal, entonces, los puntos P-P proporciones acumuladas empíricas (eje horizontal) y teóricas (eje vertical), deben estar aproximadamente en la recta de la diagonal principal. Esta es la **gráfica P-P normal** (ver figura 11.2). En este ejemplo, se puede observar que la gráfica de los porcentajes observados y esperados están aproximadamente en la diagonal principal, tal como se muestra en la figura que sigue. Sin embargo, muchas veces el determinar si la gráfica de los datos se aproxima a una recta es muy subjetivo.

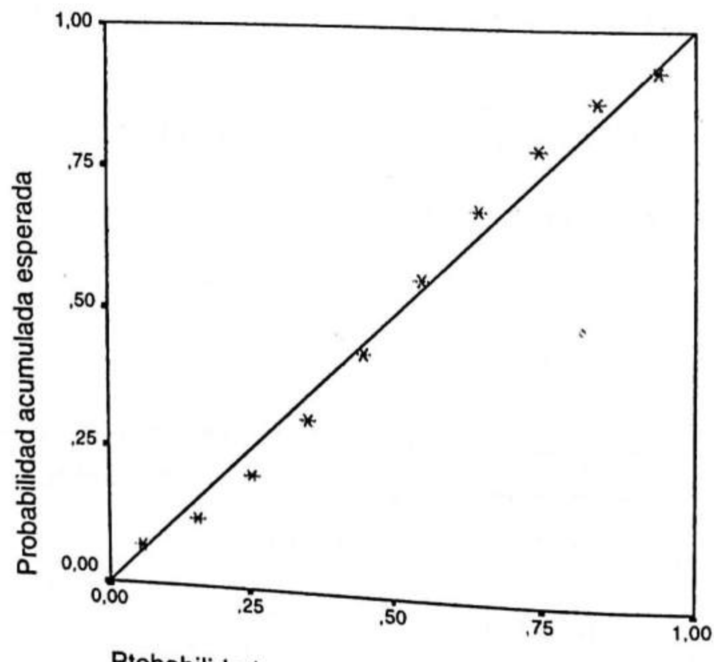


Figura 11.2. *Gráfica P-P normal*

4. El **método de Kolmogorov-Smirnov** es un procedimiento no paramétrico que se utiliza para comprobar la hipótesis nula de que la muestra procede de una población en la que la variable está distribuida según la **normal** (uniforme o Poisson). Para tomar la decisión, se debe llegar a calcular la *máxima desviación*:

$$D = \text{máxima} \mid F_0(x_i) - S_n(x_i) \mid .$$

En la tabla de una muestra de Kolmogorov-Smirnov se encuentran ciertos valores críticos de la distribución muestral de D para diversos valores de n y α. Se rechaza que los datos provienen de una distribución normal si el valor de D es mayor que el valor crítico correspondiente.

En este **ejemplo**, la diferencia máxima en valor absoluto es 0.0967. Por otra parte en la tabla de valores de Kolmogorov-Smirnov para $\alpha = 0.01$, y $n = 10$ se encuentra el valor crítico 0.490. Dado que $0.0967 < 0.490$, deberíamos aceptar la hipótesis nula donde se afirma que es normal la población de la cual se ha obtenido la muestra.

11.2 Tablas de contingencia y pruebas chi-cuadrado.

Existen 3 técnicas eficaces para el estudio de las relaciones entre dos o más variables. Estas son, *las tablas de contingencia*, *el análisis de varianza*, y el *análisis de regresión*. Básicamente una *tabla de contingencia* se obtiene al registrar los datos observados de la muestra aleatoria en doble clasificación.

Una tabla de contingencia de r filas (o renglones) y c columnas, denominada también *tabla de contingencia de dimensión r×c*, contiene en cada *entrada* o *celda* la frecuencia observada de la muestra que corresponde a dos variables clasificadas por categorías. Los totales de los renglones y los totales de las columnas se llaman *frecuencias marginales*. La suma de las frecuencias marginales es el *total n* de la muestra.

En las pruebas de hipótesis con tablas de contingencia, no se hace ninguna suposición acerca de la distribución de probabilidades de los datos.

Con las tablas de contingencia se pueden ejecutar las siguientes pruebas de hipótesis:

* *Prueba de independencia de dos variables estadísticas,*
* *Prueba de homogeneidad de muestras,* y
* *Prueba de igualdad de dos o más proporciones poblacionales.*

11.2.1. Prueba de independencia.

El procedimiento de la prueba chi-cuadrado para la bondad de ajuste de la sección anterior se puede utilizar también para probar la independencia de dos variables categóricas.

En la prueba de bondad de ajuste se tiene una sola variable estadística. Los datos se clasifican según un sólo criterio en clases o categorías y lo que se prueba es la distribución de probabilidad teórica de la variable.

Las pruebas de hipótesis de independencia implican dos variables categóricas y lo que se prueba es la suposición de que las dos variables son estadísticamente independientes.

Para cada frecuencia observada en una celda hay una frecuencia esperada que se calcula a partir de la hipótesis nula especificada y que se supone verdadera.

La Prueba

La hipótesis nula H_0, consiste en suponer que las dos variables categóricas son independientes o que los métodos de clasificación de filas y de columnas son independientes.

Si la hipótesis nula es verdadera es decir si hay independencia entre la variable de la fila y la variable de la columna, entonces, la probabilidad de una observación en cada una de las rc celdas es estimada por:

$$p = P[\text{fila} = i, \text{columna} = j] = \frac{\text{total de fila } i}{n} \times \frac{\text{total de columna } j}{n}$$

La frecuencia esperada correspondiente en cada una de las rc celdas se obtiene multiplicando la probabilidad por el total n de la muestra, esto es,

$$\textit{Frecuencia esperada} = np = \frac{\text{total de fila} \times \text{total de columna}}{\text{Gran total}}$$

La estadística para la prueba de independencia es la misma que se utiliza en bondad de ajuste de la sección anterior. El número de grados de libertad está dado por:

$$\nu = (r-1)(c-1).$$

Dado un nivel de significación α, en la tabla chi-cuadrado se encuentra el valor crítico que denotaremos por $\chi^2_{1-\alpha,\nu}$.

Para probar la hipótesis nula de independencia se utiliza el siguiente criterio de decisión: Calcular la cantidad:

$$\chi^2_{cal} = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

en donde, la suma se extiende a todas las rc celdas en la tabla de contingencia $r \times c$. Si $\chi^2_{cal} > \chi^2_{1-\alpha,\nu}$, se rechaza H_0, en caso contrario se aceptará H_0.

NOTA. Cuando la muestra es pequeña, digamos menor de 50, o cuando algunas o todas las frecuencias de las celdas son menores que 5, o cuando el grado de libertad es igual a 1, debe aplicarse la corrección de Yates. El cálculo de chi-cuadrado corregida por Yates se efectúa por:

$$\chi^2_{cal} = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

NOTA. (Coeficiente de contingencia). Es una medida de la correlación entre dos variables categóricas cuyos valores se registran en una tabla de contingencia.
El coeficiente de contingencia se define por:

$$C = \sqrt{\frac{\chi^2_{cal}}{n + \chi^2_{cal}}}$$

en donde, n es el tamaño de la muestra.
Si χ^2_{cal} es significativo, también lo es el coeficiente de contingencia.

EJEMPLO 11.6.

Una socióloga quiere determinar si hay alguna relación entre el tamaño de la familia y el nivel de educación del padre. Para esto, escogió una muestra de 500 hogares y los clasificó de acuerdo con dos criterios: por el tamaño de la familia y por el nivel de educación. Las frecuencias observadas están registradas en tabla 11.6. ¿Se puede concluir al nivel de significación de 0.05 que el tamaño de la familia es independiente del nivel de educación del padre?.

Tabla 11.6. *500 hogares clasificados por tamaño de familia y número de hijos*

	Número de hijos		
Nivel de educación	Menos de 3	[3, 5]	Más de 5
Primaria	40	90	70
Secundaria	50	60	60
Superior	60	50	20

SOLUCION.

1. *Hipótesis*:
   H_0 : El tamaño de la familia es independiente del nivel de educación del padre..
   H_1 : El tamaño de la familia depende del nivel de educación del padre.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística.* $\sum_i \frac{(O_i - e_i)^2}{e_i}$, se distribuye aproximadamente como chi-cuadrado

   con $\nu = (r-1)(c-1) = (3-1)(3-1) = 4$ grados de libertad.
4. *Región crítica*: Para el nivel de significación $\alpha = 0.05$ y 4 grados de libertad el valor crítico es: $\chi^2_{0.95,4} = 9.49$. Se rechazará H_0 si el valor calculado de la estadística chi-cuadrado es mayor que 9.49.
5. *Cálculos*: Sólo se necesitan calcular las frecuencias esperadas de 4 celdas las otras frecuencias esperadas se encuentran por sustracción debido a que la suma de las frecuencias esperadas en cualquier fila o columna debe dar el total marginal. Las frecuencias observadas y las esperadas (en paréntesis) se dan en la tabla 11.7:

Tabla 11.7. Frecuencias observadas y esperadas

Nivel de Educación	Número de hijos			Total
	Menos de 3	[3, 5]	Más de 5	
Primaria	40 (60)	90 (80)	70 (60)	200
Secundaria	50 (51)	60 (68)	60 (51)	170
Superior	60 (39)	50 (52)	20 (39)	130
Total	160	200	150	500

Luego,

$$\chi^2_{cal} = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(40-60)^2}{60} + \frac{(90-80)^2}{80} + \cdots + \frac{(20-39)^2}{39} = 3.2.77$$

6. *Decisión*: Dado que $32.77 < 9.49$, se debe rechazar, H_0 .

NOTA. La probabilidad *P* es: $P = P[\chi^2(4) > 9.49] = 0.000$.

NOTA. El **coeficiente de contingencia** de las dos variables de la tabla es:

$$C = \sqrt{\frac{\chi^2_{cal}}{n + \chi^2_{cal}}} = \sqrt{\frac{3.2.77}{500 + 332.77}} = 0.248$$

Dado que $\chi^2_{cal} = 32.77$ es significativo, entonces, C=0.248 indica que las dos variables están correlacionadas.

11.2.2. Prueba de homogeneidad de muestras

El método de prueba de independencia es aplicable también a la prueba de homogeneidad de muestras.

Con la prueba de homogeneidad de muestras se busca determinar si dos o más muestras independientes provienen de una misma población.

Como en el método anterior, para esta prueba los datos muestrales se registran en *rc* celdas de una tabla de contingencia de orden $r \times c$.

Las hipótesis nula y alternativa de la prueba de homogeneidad son respectivamente:

H_0: Las muestras aleatorias provienen de una misma población. (o las muestras son homogéneas).

H_1: Las muestras aleatorias no provienen de una misma población (o las muestras no son homogéneas).

El proceso de la prueba de estas hipótesis es el mismo de la sección anterior.

EJEMPLO 11.7.

El departamento medico de la PUCP estudia el nivel de efectividad de tres remedios para la alergia: R1, R2 y R3. Cada remedio se suministró a 60 profesores. Los resultados del experimento se dan en la tabla que sigue.

	Remedios para la alergia		
Efectividad	R1	R2	R3
Sin alivio	10	20	15
Cierto alivio	40	30	20
Alivio total	10	10	25

¿Podemos concluir al nivel de significación 0.01, que los tres remedios para la alergia son igualmente efectivos?

SOLUCION.

1. *Hipótesis*:
   H_0: Los tres remedios para la alergia, R1, R2 y R3 son igualmente efectivos.
   H_1: Los tres remedios para la alergia no son igualmente efectivos.
2. *Nivel de significación*: 0.01.
3. *Estadística.* $\sum_i \frac{(O_i - e_i)^2}{e_i}$, que se distribuye aproximadamente como chi-cuadrado, con grados de libertad igual a $\nu = (r-1)(c-1) = (3-1)(3-1) = 4$

4. *Región crítica*: Para el nivel de significación 0.01 y 4 grados de libertad el valor crítico es: $\chi^2_{0.99,4} = 13.28$.

   Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor que 13.28. En caso contrario, no se debería rechazar H_0.

5. *Cálculos*: Las frecuencias observadas y las esperadas (en paréntesis) se dan en la tabla 11.8

Tabla 11.8. *Frecuencias observadas y esperadas()*

Efectividad	Remedios para la alergia			Total
	R1	R2	R3	
Sin alivio	10 (15)	20 (15)	15 (15)	45
Cierto alivio	40 (30)	30 (30)	20 (30)	90
Alivio total	10 (15)	10 (15)	25 (15)	45
Total	60	60	60	180

Luego,

$$\chi^2_{cal} = \frac{(10-15)^2}{15} + \frac{(20-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(40-30)^2}{30} + \frac{(30-30)^2}{30}$$

$$+ \frac{(20-30)^2}{30} + \frac{(10-15)^2}{15} + \frac{(10-15)^2}{15} + \frac{(25-15)^2}{15} = 20$$

5. Dado que 20 >13.28, se debería rechazar la hipótesis nula H_0 al nivel de significación de 0.01.

NOTA. La probabilidad P de la prueba es: $P=P[\chi^2(4)>20]=0.0005$.

11.2.3. Prueba para más de dos proporciones

Un caso particular de la prueba de homogeneidad de muestras, es la prueba de la hipótesis nula que establece que los k $(k \geq 2)$ parámetros binomiales (o porcentajes de éxitos poblacionales) tienen el mismo valor.

Para esta prueba, los tamaños de las muestras y el número de éxitos en cada una de las muestras se ordenan como en la tabla 11.9,

Tabla 11.9. *k muestras independientes*

	Muestras			
	1	2	...	k
Exitos	x_1	x_2	...	x_k
Fracasos	$n_1 - x_1$	$n_2 - x_2$		$n_k - x_k$
Total	n_1	n_2	...	n_k

En la tabla 11.9, k es el número de muestras aleatorias independientes, obtenidas respectivamente de k poblaciones binomiales cuyos parámetros respectivos son: $p_1, p_2, ..., p_k$. Además, $n_1, n_2, ..., n_k$ son los tamaños de las k muestras, y $x_1, x_2, ..., x_k$, representan el número de éxitos en la muestra respectiva .

La hipótesis nula consiste pues en afirmar que todas las proporciones poblacionales son iguales. Esto es,

$$H_0: \ p_1 = p_2 = \cdots = p_k$$

La hipótesis alternativa consiste en afirmar que no todas las proporciones poblacionales son iguales.

EJEMPLO 11.8 .

En una fábrica de confecciones textiles que opera en tres turnos; mañana, tarde y noche, el jefe de control de calidad, quiere saber si hay diferencias en la calidad de las confecciones en los tres turnos. Para esto tomó una muestra aleatoria de 100 confecciones de cada turno del día anterior y las clasificó según el turno de su producción y según su calidad en defectuoso o no defectuoso. Los resultados se dan en la siguiente tabla:

	Turnos		
Calidad	Mañana	Tarde	Noche
Defectuosos	3	12	15
No defectuosos	97	88	85

¿Podemos concluir al nivel de significación del 5%, que la proporción de todas las confecciones defectuosas es la misma para los tres turnos?

SOLUCION.

Sean p_1, p_2, p_3 los porcentajes de confecciones defectuosas para los tres turnos: mañana, tarde y noche, respectivamente.

1. *Hipótesis*:
   H_0: $p_1 = p_2 = p_3$
   H_1: No todos los p_1, p_2, p_3 son iguales
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*. $\sum_i \frac{(O_i - e_i)^2}{e_i}$, se distribuye aproximadamente como chi-cuadrado con $\nu = (r-1)(c-1) = (2-1)(3-1) = 2$ grados de libertad.
4. *Región crítica*: Para el nivel de significación $\alpha = 0.05$ y 2 grados de libertad, el valor crítico $\chi^2_{0.95,\,2} = 5.99$.
   Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor de 5.99. No se rechazará en caso contrario.
5. *Cálculos*: Las frecuencias observadas y esperadas (en paréntesis) se dan en la tabla que sigue:

Tabla 11.10. *Frecuencias observadas y esperadas*

Calidad	Muestras			
	Mañana	Tarde	Noche	Total
Defectuosos	3 (10)	12 (10)	15 (10)	30
No defectuosos	97 (90)	88 (90)	85 (90)	270
Total	100	100	100	300

Entonces,

$$\chi^2_{cal} = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(3-10)^2}{10} + \frac{(12-10)^2}{10} + \cdots + \frac{(85-90)^2}{90} = 8.667$$

6. *Decisión*: Dado que $8.667 > 5.99$, se debería rechazar, H_0.

NOTA. La significación de la prueba es: $P = P[\chi^2(2) > 8.667] = 0.0131$. Observe que no se deberá rechazar H_0 al nivel de significación 1%.

EJERCICIOS

1. Un jugador quiere probar que es legal el dado con el que juega. Tiró el dado 120 veces y obtuvo la siguiente distribución de frecuencias de las caras resultantes

Resultado	1	2	3	4	5	6
Frecuencia	15	25	33	17	16	14

a) Enuncie las hipótesis de la prueba y determine las frecuencias esperadas
b) Describa la estadística de la prueba
c) Determine la región crítica de la prueba al nivel de significación del 5%.
d) ¿A que conclusión llega usando el nivel de significación 0.05?
e) Determine la probabilidad *P*

Rp. a) H_0: la distribución de probabilidad es uniforme, frec. esperadas: 20, 20, 20, 20, 20, 20, b) $\chi_k^2 = 14$, gl. = 5, c) $R.C. =]15.09, +\infty[$, d) se acepta H_0. , e) Significación: 0.02.

2. El gerente de ventas de una la compañía P&C afirma que todos sus vendedores realizan el mismo número de visitas durante el mismo periodo de tiempo. Una muestra aleatoria de 5 registros de los vendedores en una semana dada reveló el siguiente número de visitas.

Vendedor	A	B	C	D	E
Número de vistas	23	29	25	23	30

Con el nivel de significación de 0.05, ¿es razonable aceptar la afirmación del gerente?

Rp. H_0: realizan el mismo número de visitas, frec. esperadas: 26, 26, 26, 26, 26, $\chi_k^2 = 1.69$, gl =4 , $R.C. =]9.49, +\infty[$, se acepta H_0. (Significación: 0.792)

3. El gerente de personal de la compañía "REXA" quiere probar la hipótesis que hay diferencias significativas de tardanzas de los diferentes días de la semana. De los registros de asistencia obtuvo la siguiente tabla de tardanzas de su personal para cada uno de los días de la semana:

Días	Lunes	Martes	Miércoles	Jueves	Viernes
Tardanzas	58	39	75	48	80

¿Se puede aceptar la hipótesis del gerente con un nivel de significación de 0.05?

Rp. H_0: el número de tardanzas es el mismo cada día, frec. esperadas: 60, 60, 60, 60, 60, $\chi_k^2 = 20.233$, gl =4 , $R.C. =]9.49, +\infty[$, se rechaza H_0. (Significación: 0.000)

4. De una muestra de turistas que se hospedan en el hotel "EL PALMER" se recogió sus opiniones acerca de los servicios del hotel, resultando los siguientes datos:

Pésima	Mala	Regular	Buena	Muy buena	Excelente
Turistas	20	25	40	54	56

Pruebe con un nivel de significación del 5%, la hipótesis nula de que no hay diferencias significativas entre las opiniones de los turistas.

Rp. H_0: No hay diferencias en las opiniones, frec. esperadas: 39, 39, 39, 39, 39, $\chi_k^2 = 27.487$, gl =4 , $R.C. =]9.49, +\infty[$, se rechaza H_0. (Significación: 0.000)

5. En un día dado se observó el número de conductores que escogieron cada una de las 10 casetas de pago de peaje ubicadas a la salida al sur. Los datos se registraron en la siguiente tabla:

Caseta #	1	2	3	4	5	6	7	8	9	10
# de conductores	580	700	730	745	720	710	660	655	670	490

¿Presentan estos datos suficiente evidencia para concluir que hay casetas preferidas?. Utilice el nivel de significación del 5%.

Rp. H_0 : No hay casetas preferidas, frec. esperadas: 666, 666, 666, 666, 666, 666, 666, 666, 666, 666, $\chi_k^2 = 82.42$, gl. = 9, $R.C. =]16.92, +\infty[$, se rechaza H_0.

6. Un ejecutivo del hipermercado "TOD" afirma que los compras se pagan 30% con cheques, 45% con efectivo y 25% con tarjeta de crédito. En una muestra aleatoria de 400 compradores se encontró que 110 de ellos pagaron con cheques, 210 con efectivo y 80 con tarjetas. ¿Puede usted concluir, con la significación de 0.05, que la afirmación del ejecutivo es razonable?.

Rp. H_0: Los pagos con cheque, con efectivo, y con tarjeta de crédito, guardan la relación : 30:45:25, frec. esperadas: 120, 180, 100, $\chi_k^2 = 9.833$, gl =2 , $R.C. =]5.99, +\infty[$, se rechaza H_0

7. Una máquina llena latas con 300 caramelos de sabores: Piña fresa, limón y naranja en la relación: 4:3:2:1. Si en una lata de estos caramelos se encontró: 115 de piña, 95 de fresa, 70 de limón y 20 de naranja, pruebe la hipótesis de que la máquina está mezclando en la relación: 4:3:2:1, al nivel de significación de 0.05.

Rp. H_0: la máquina está mezclando en la relación: 4:3:2:1, frec. esperadas: 120, 90, 60, 30. $\chi_k^2 = 5.496$, gl =3 , $R.C. =]7.82, +\infty[$, se acepta H_0.

8. Se cree que las personas que mueren por sobredosis de narcóticos son generalmente jóvenes. Para comprobar esta hipótesis se ha obtenido la siguiente distribución del número de muertes por sobredosis.

Edad	15 –19	20 –24	25 –29	30 –34	35 –39	40 o más
Número de muertes	31	44	27	39	41	28

Con estos resultados y con un nivel de significación de 0.05, ¿se puede concluir, empleando, que muere un número igual de personas en cada categoría?.

Rp. H_0: muere igual número de personas en cada categoría, frec. esperadas: 35, 35, 35, 35, 35, 35,

$\chi_k^2 = 7.486$, gl =5, $R.C. =]11.07, +\infty[$, no se debe rechazar H_0. (Significación: 0.187)

9. Un investigador escogió una muestra aleatoria de 192 familias con 4 hijos y encontró la siguiente distribución de frecuencias del número de hijos varones:

Número de varones	0	1	2	3	4
Número de familias	18	42	64	40	28

El quiere probar la hipótesis de que los nacimientos de varones y mujeres son igualmente probables. Esto es, quiere probar que la distribución de estos datos se aproxima a una distribución binomial..

a) Enuncie las hipótesis de la prueba y obtenga las frecuencias esperadas
b) Describa la estadística de la prueba
c) Determine la región crítica de la prueba al nivel de significación del 5%.
d) ¿A que conclusión llega usando el nivel de significación 0.05?
e) Determine el nivel de significación de la prueba. (Calcule probabilidad: P)

Rp. a) H_0: nacimientos de varones es B(4,0.5), frec. esperadas: 9.9, 43.5, 71.7, 52.5, 14.4,

b) $\chi_k^2 = 23.28$, gl =4 , c) $R.C. =]13.28, +\infty[$, d) se rechaza H_0. e) Significación: 0.000

10. Se lanzaron 200 veces 5 monedas y en cada tirada se contaron el número de caras. Los resultados de este experimento son los siguientes:

Número de caras	0	1	2	3	4	5
Número de tiradas	3	15	55	60	40	27

Pruebe la hipótesis de que la distribución del número de caras se ajusta a una distribución binomial. Use el nivel de significación del 1%.

Rp. $6\hat{p} = \bar{x} = 3$, $\hat{p} = 0.6$, H_0: la distribución es binomial B(5, 0.6), frec. esperadas: 2.1, 15.4, 46.1, 69.1, 51.8, 15.6, $\chi_k^2 = 14.08$, gl. = 3, $R.C. =]11.35, +\infty[$, se rechaza H_0. (Signific: 0.003)

11. Un vendedor de la compañía "ELECTRIC" visita a 6 clientes por día. Se cree que el número de ventas por día que él realiza es una variable aleatoria que puede ser descrita mediante una distribución Binomial. Durante 100 días se han registrado las siguientes ventas por día de este vendedor.

Número de ventas	0	1	2	3	4	5
Número de días	10	41	60	20	6	3

En el nivel de significación 0.01, ¿puede usted concluir que estos datos, en efecto, se ajustan a la distribución binomial B(5, 0.4)?.

Rp. H_0: La distribuc es $B(5,0.4)$, frec. esperadas: 11, 36, 48, 32, 11, 1, $\chi_k^2 = 32.667$, gl. = 3, $R.C. =]7.82, +\infty[$, se rechaza H_0. (Significación: 0.000)

12. Se seleccionaron aleatoriamente 100 cuentas en la sección de contabilidad del banco "CREDIT" y se las examinó para descubrir errores, obteniendo los siguientes resultados.

Número de errores	0	1	2	3	4	5	6
Número de cuentas	64	46	25	9	4	1	1

Pruebe la hipótesis de que la distribución del número de errores se ajusta a una distribución de Poisson con media uno. Use un nivel de significación del 1%.

Rp. $\hat{\lambda} = \bar{x} = 1$, H_0: la distribución es de Poisson $P(1)$, frec. esperadas: 55.2, 55.2, 27.6, 9.2, 2.3, 0.5, 0.1, $\chi_k^2 = 3.91$, gl. = 2, $R.C. =]9.21, +\infty[$, se acepta H_0. (Significación: 0.141)

13. Durante 100 intervalos de tiempo cada uno de 3 minutos, se registraron las llamadas telefónicas recibidas en una central, resultando la siguiente distribución:

Número de llamadas	0	1	2	3	4
Número de intervalos	48	35	11	5	1

¿Puede usted concluir, con probabilidad de error tipo I de 0.01, que la distribución de estos datos se ajusta a una distribución de Poisson?.

Rp. $\hat{\lambda} = \bar{x} = 0.76$, H_0: la distribución es de Poisson $P(0.76)$, frec. esperadas: 46.8, 35.5, 13.5, 3.4, 0.7, $\chi_k^2 = 1.42$, gl. = 2, $R.C. =]9.21, +\infty[$, se acepta H_0. (Significación: 0.492)

14. El servicio médico de la universidad trata de determinar si el número de pacientes que atiende por emergencia tiene distribución de Poisson. Durante 100 días se han registrado el número de pacientes que son atendidos por emergencia resultando la siguiente distribución de frecuencias:

Número de pacientes	0	1	2	3	4	5
Número de días	40	34	16	7	2	1

Al nivel de significación 0.05, ¿se puede concluir que los datos, en efecto, se ajustan a la distribución sugerida?.

Rp. $\hat{\lambda} = \bar{x} = 1$, H_0: la distribución es de Poisson $P(1)$, frec. Esperadas (mayores 1ue 5): 36.8, 36.8, 18.4, 8, $\chi_k^2 = 1.312$, gl. = 2, *R.C.*=]5.99, +∞[, se acepta H_0. (Significación: 0.51687)

15. A menudo se dice que los profesores tienden a clasificar a sus alumnos de acuerdo con la *curva normal*. En un examen final del curso de estadística general las calificaciones de los alumnos, han sido tabuladas en la distribución de frecuencias que sigue.

Notas	Alumnos
[02, 05[	4
[05, 08[	16
[08, 11[	55
[11, 14[	47
[14, 17[	28
[17, 20]	10

¿Podemos concluir, con un nivel de significación de 0.01, que esta distribución de notas concuerda con una distribución normal?.

Rp. H_0: la distribución es normal con $\bar{x} = 11.544$, $\hat{s} = 3.421$, frec. esperadas: 4.5, 19.6, 45.9, 52.3, 28.97, 7.8, $\chi_k^2 = 3.69$, gl. = 2, *R.C.* =]9.21, +∞[, se acepta H_0. (Significación: 0.158).

16. En una muestra aleatoria de 45 baterías de la marca "BOCH" para automóvil, vendidas hace dos años, se encontraron los siguientes tiempos de vida útil en meses:

Vida útil	Baterías
[10.5, 11.5[	2
[11.5, 12.5[	4
[12.5, 13.5[	6
[13.5, 14.5[	10
[14.5, 15.5[	15
[15.5, 16.5[	5
[16.5, 17.5]	3

¿Se puede concluir, con un nivel de significación de 0.05, que la duración de tales baterías tiene una distribución normal con media 14.311 horas y desviación estándar de 1.4897?

Rp. H_0: la distribución es normal con $\bar{x} = 14.311$, $\hat{s} = 1.4897$, frec. esperadas:1.3, 3.7, 8.1, 11.6, 10.7, 6.4, 2.5, $\chi_k^2 = 2.799$, gl. = 2, *R.C.* =]5.99, +∞[, se acepta H_0. (Significación: 0.247).

17. Los ingresos mensuales en dólares de una muestra de hogares de la ciudad de Tarapoto se tabularon en la siguiente distribución de frecuencias:

Ingresos	Familias
[150, 200[	20
[200, 250[	30
[250, 300[	60
[300, 350[	100
[350, 400[	80
[400, 450[	60
[450, 500]	40

¿Puede usted concluir, con el nivel de significación del 1%, que estos ingresos familiares tienen una distribución normal?.

Rp. H_0: la distribución es normal con $\overline{x} = 342.499$, $\hat{s} = 97.755$, frec. esperadas: 14.3, 33.3, 67.6, 93.6, 88.7, 57.5, 25.5, $\chi_k^2 = 13.19$, gl. = 4, *R.C.* =]13.28, +∞[, se acepta H_0.

18. Al nivel de significación del 1%, ¿podemos concluir que es normal $N(0, 1)$ la población de la que ha sido extraída la muestra aleatoria simple:

2.6, 1.8, 2.2, −1.6, 2.1, 2.7, −1.0, 2.0, 1.7, 2.4?.

Rp. H_0: la distribución es normal N(0,1). La diferencia máxima de porcentajes observados y esperados es D= 0.6554, para $n = 10$, $\alpha = 0.01$, en la tabla de Kolmogorov-Smirnov se halla el valor crítico 0.490. Dado que 0.6554 > 0.490 se rechaza la hipótesis nula.

19. ¿Podemos concluir que es normal la población de la que ha sido extraída la siguiente muestra aleatoria simple:

−8.01, 11.53, 10.8, −3.14, −13.9, 8.63, −1.79, −5.63, 5.04, 14.36 ?.

Use el nivel de significación: 0.01.

Rp. H_0: la distribución es normal con $\overline{x} = 1.789$, $\hat{s} = 9.5789$. La diferencia máxima de porcentajes observados y esperados es $D = 0.162$, para $n = 10$, $\alpha = 0.01$, en la tabla de Kolmogorov-Smirnov se halla el valor crítico 0.490. Dado que 0.162 < 0.490 se acepta la hipótesis nula.

20. Las calificaciones de 350 personas sometidas a una prueba de aptitud dieron la siguiente tabla de frecuencias:

Calificación	80	85	90	95	100	105	110	115	120
Personas	13	27	38	55	70	60	42	33	12

¿Puede usted concluir, con el nivel de significación de 0.05, que estos datos se ajustan a una distribución normal?.

Rp. H_0: la distribución es normal con $\overline{x} = 100.3857$, $\hat{s} = 9.917$, $D = 0.01763$, para $n = 350$. α=0.05, en la tabla de K-S se halla el valor crítico $1.36/\sqrt{350}$ =0.0727. Dado que 0.01763<0.0727 se acepta la hipótesis nula

21. El gerente de procesamientos de datos de la compañía "COMP" estudia el uso de la computadora en el departamento de contabilidad de la compañía. En una muestra aleatoria de 60 trabajos del mes pasado se registró el tiempo de procesamiento (en segundos) para cada trabajo, con los siguientes resultados:

03	16	18	04	15	12	17	11	15	17	07	02	08	14	16
09	10	14	18	14	12	19	13	09	07	19	11	15	17	12
16	05	09	14	16	17	14	16	10	15	14	17	11	10	15
13	16	13	15	12	09	08	15	17	09	14	13	10	15	13

Al nivel de significación del 5%, pruebe la hipótesis de que la distribución los tiempos de procesamiento es normal.

Rp. H_0: la distribución es normal con $\bar{x} = 12.75$, $\hat{s} = 3.9644$. La diferencia máxima de porcentajes observados y esperados es $D = 0.140$, para $n = 60$, $\alpha = 0.05$, en la tabla de Kolmogorov-Smirnov se halla el valor crítico $1.36 / \sqrt{60} = 0.1756$. Dado que $0.140 < 0.1755$ se acepta la hipótesis nula

22. Los siguientes resultados representan las calificaciones en una prueba de aptitud académica que rindieron 150 alumnos:

68, 68, 68, 68, 68, 68, 68, 68, 68, 68, 75, 75, 75, 76, 76, 76, 76, 56, 56, 56, 56
56, 57, 57, 57, 57, 57, 65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 72, 72, 72, 72, 72
73, 73, 73, 73, 73, 35, 36, 37, 38, 39, 69, 69, 69, 69, 69, 69, 80, 81, 81, 82, 82,
83, 83, 84, 70, 70, 70, 70, 70, 71, 71, 71, 71, 71, 58, 58, 58, 59, 59, 59, 58, 58
67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 45, 46, 46, 47, 47, 47, 48, 48, 74, 74, 74
74, 74, 70, 71, 72, 40, 41, 41, 42, 43, 43, 44, 42, 78, 78, 78, 79, 79, 50, 51, 51
52, 52, 53, 53, 53, 54, 54, 60, 61, 61, 61, 62, 62, 62, 62, 62, 62, 63, 64, 85, 86
87, 30, 93

Suponiendo que estas notas constituyen una muestra aleatoria, ¿puede usted concluir, con un nivel de significación del 5%, que estas notas provienen de una distribución normal?

Rp. H_0: la distribución es normal con $\bar{x} = 63.88$, $\hat{s} = 12.2759$. La diferencia máxima de porcentajes observados y esperados es $D = 0.123$, para $n = 150$, $\alpha = 0.01$, en la tabla de Kolmogorov-Smirnov se halla el valor crítico $1.63 / \sqrt{150} = 0.133$. Dado que $0.123 < 0.133$ se acepta la hipótesis nula

23. Con los datos del problema 21 construya una distribución de frecuencias de 13 intervalos y utilizando el método de chi-cuadrado pruebe la hipótesis nula que la población de las calificaciones es normal. Use el nivel de significación: 0.05..

Rp. Con 13 intervalos se obtiene: frec. Observadas: 1,5, 8, 8, 10, 18, 12, 36, 28, 12, 8, 3, 1, H_0: esta distribución es normal con $\bar{x} = 64.367$, $\hat{s} = 12.241$, frec. Esperadas: 1.2, 2.3, 5.03, 9.6, 15.3, 20.8, 23.99, 23.5, 19.5, 13.8, 8.2, 4.2, 1.8, $\chi_k^2 = 23.16$, gl. = 7, $R.C. =]14.07, +\infty[$, se acepta H_0. (Significación: 0.002). A la misma conclusión se llega agrupando en cualquier número de intervalos entre 5 y 20.

24. A una muestra de empleados de la PUCP clasificados como: Docentes, no docentes y de servicio, se les pidió que escogieran entre tres planes de seguro familiar particular: A B y C. En el cuadro que sigue se dan los resultados:

Clase de trabajo	Plan de seguro		
	A	B	C
Docente	100	150	60
No docente	40	70	20
Servicios	20	40	10

Se quiere probar si hay relación entre el plan de seguro que seleccionaron y su tipo de trabajo.

a) Enuncie las hipótesis de la prueba.
b) Describa la estadística de la prueba
c) Determine la región crítica de la prueba al nivel de significación del 5%.
d) ¿A que conclusión llega usando el nivel de significación 0.05?
e) Determine el nivel de significación de la prueba. (Calcule probabilidad: P)

Rp. a) H_0: El tipo de plan no depende del tipo de trabajo, b) $\chi_k^2 = 2.745$, gl.=4, c) $R.C. =]9.49, +\infty[$, d) se acepta H_0. e) Significación: 0.601.

25. Una muestra aleatoria de 200 adultos se clasificó de acuerdo a su sexo y al número de horas que miran televisión durante la semana. Las frecuencias observadas se dan en la siguiente tabla.

Sexo	Número de horas que miran T.V.	
	Menos de 15 horas	al menos 15 horas
Hombres	55	45
Mujeres	40	60

Con esta información, ¿se puede concluir con un nivel de significación de 0.05, que el tiempo utilizado para ver televisión es independiente del sexo?.

Rp. H_0: El número de horas que miran T.V. es independiente del sexo, $\chi_k^2 = 4.51$, gl. = 1, $R.C. =]3.84, +\infty[$, se rechaza H_0. (Significación: 0.034)

26. Se seleccionó una muestra de 800 votantes y se les clasificó de acuerdo a su nivel de ingresos como; bajo, medio, alto, y según su opinión con respecto a una reforma impositiva en: a favor, en contra, sin decisión. Las frecuencias observadas se dan en la siguiente tabla.

Opinión	Ingresos		
	Bajo	Medio	Alto
A favor	200	130	70
En contra	60	60	80
Sin decisión	40	60	100

¿Hay relación entre la opinión de los votantes y su nivel de ingresos?. Use el nivel de significación 0.05.

Rp. H_0: la opinión no depende los ingresos, $\chi_k^2 = 88$, gl. = 4, $R.C. =]9.49, +\infty[$, se rechaza H_0. (Significación: 0.000).

27. Un investigador realizó un estudio para determinar si el tamaño de familia depende del nivel de educación del padre. La muestra se clasificó de acuerdo al nivel de educación y al número de hijos, en la siguiente tabla.

Nivel de educación	Número de hijos				
	0 o 1	2	3	4	5 o más
Primaria	20	18	12	14	30
Secundaria	50	25	18	16	24
Superior	12	6	4	8	12

Con estos datos, ¿se puede inferir que el tamaño de la familia es independiente del nivel de educación del padre?. Use el nivel de significación 5%.

Rp. H_0: El tamaño de familia es independiente de la educación del padre, $\chi_k^2 = 11.525$, gl. = 8, $R.C. =]15.51, +\infty[$, se acepta H_0. (Significación: 0.174).

28. El gerente de ventas de la firma "GROUP" desea determinar si las ventas de cuatro productos dependen de la clase de clientes clasificados en cuatro grupos. Una muestra aleatoria de las ventas suministró la siguiente información:

Grupo de clientes	Producto			
	1	2	3	4
Profesionales	30	35	55	40
Comerciantes	155	50	125	80
Obreros	130	30	105	50
Amas de casa	35	15	20	45

¿Se puede concluir que las ventas de los 4 productos es homogénea entre los 4 grupos de clientes?. Use el nivel de significación 0.05.

Rp. H_0: las ventas de los 4 productos es homogénea entre los 4 clases de clientes, $\chi_k^2 = 61.07$, gl.=9, $R.C. =]16.92, +\infty[$, se rechaza H_0. (Significación: 0.000)

29. El cuadro que sigue es una muestra de televidentes por clases sociales que ven diariamente cuatro programas de T.V. del mediodía (talk-Shows):
Enuncie la hipótesis H_0 para determinar si hay independencia estadística entre los televidentes de talk-shows y clases sociales, y probar tal hipótesis al nivel de significación al 1%.

Talk-shows	Clase social			
	Pobre	Media baja	Media	Rica
TS1	190	280	500	280
TS2	250	300	350	150
TS3	160	250	180	120
TS4	100	150	80	80

Rp. H_0: la preferencia del programa es independiente de la clase social. $\chi_k^2 = 139.977$, gl. = 9, $R.C. =]21.67, +\infty[$, se rechaza H_0. (Significación: 0.000).

30. La compañía de cerveza "DORADA" está interesada en saber si el consumo de su marca de cerveza depende de una localización geográfica en especial. Para responder a esta pregunta la compañía solicitó a una firma de investigación de mercados interrogar a consumidores en cada una de las cuatro regiones principales del país. Los resultados se registran en la siguiente tabla

Consumo anual Por persona	Región			
	Sur	Centro	Norte	Selva
Más de 10 cajas	200	120	100	90
De 5 a 10 cajas	150	100	150	200
Menos de 5 cajas	100	230	200	160

Al nivel de significación del 5%, ¿la localización geográfica es un factor insignificante en el consumo de la cerveza?.

Rp. Frecuencias. Esperadas por filas: 127.5, 127.5, 127.5, 150, 150, 150, 172.5, 172.5, 172.5. $\chi_k^2 = 146.89$, gl. = 6, $R.C. =]12.59, +\infty[$, se rechaza H_0. (Significación: 0.000).

31. El cuadro que sigue es una muestra de lectores por clases sociales que leen diariamente cuatro periódicos:

: Periódico	Clase social			
	Pobre	Clase media inferior	Clase media	Rica
X	14	25	45	24
Y	20	25	30	10
Z	11	20	14	4
W	10	15	10	6

Enuncie las hipótesis para determinar si hay relación entre lectores de periódicos y clases sociales. Realice la prueba al nivel de significación 0.01.

Rp. H_0 : la preferencia de los periódicos es independiente de la clase social, $\chi_k^2 = 16.896$, gl. = 9, $R.C. =]21.67, +\infty[$, se acepta H_0. (Significación: 0.05037).

32. El director de compras de una fábrica grande debe decidir por la compra de una de cuatro marcas de máquinas que hay en el mercado. Para probar si existe diferencia significativa en la calidad de las máquinas, obtiene una muestra de la producción de 150 artículos para cada una de ellas y observa el número de defectuosos. Los resultados se dan en la siguiente tabla:

Calidad	Máquinas			
	A	B	C	D
Defectuosos	21	12	15	18
Buenos	129	138	135	132

a) Enuncie las hipótesis de la prueba.
b) Describa la estadística de la prueba
c) Determine la región crítica de la prueba al nivel de significación del 5%.
d) ¿A que conclusión llega usted si usa el nivel de significación 0.05?
e) Determine la probabilidad P

Rp. a) H_0: las cuatro proporciones son iguales, b) $\chi_k^2 = 3.06435$, gl. = 3, c) $R.C. =]7.82, +\infty[$, se acepta H_0. e) Significación: 0.3818.

33. Se realiza un estudio para determinar si existe diferencias significativas entre las proporciones de adultos de las ciudades de Lima, Cuzco, Trujillo e Iquitos que prefieren una determinada pasta dental. Las respuestas de 200 adultos seleccionados al azar en cada una de estas ciudades se registran en la siguiente tabla:

Prefieren	Ciudad			
	Lima	Cuzco	Trujillo	Iquitos
Si	130	125	135	140
No	70	75	65	60

En el nivel de significación de 0.05, ¿se puede inferir que las proporciones de personas que prefieren la pasta dental son las mismas en las 4 ciudades?.

Rp. H_0.: Las 4 proporciones de preferencias son iguales: $\chi_k^2 = 2.795$, gl.=3, $R.C. =]9.35, +\infty[$, se acepta H_0. (Significación: 0.424).

34. El departamento de reclamos de la compañía "TP&TP" cree que la actitud de reclamar de sus clientes es independiente de la edad. Se quiere verificar esta hipótesis utilizando una muestra aleatoria de 600 clientes cuyas respuestas se dan en la siguiente tabla de contingencia:

Reclama	Grupo de edad			
	De 16 a 25	De 25 a 40	De 40 a55	55 o mayor
No	135	140	100	120
Si	15	40	20	30

Al nivel de significación del 5%,

a) ¿Cuál es su opinión respecto a la hipótesis propuesta?

b) ¿Puede usted concluir que las proporciones de clientes que reclaman es la misma en las cuatro categorías de edades?.

c) Utilice la estadística Z para determinar si la proporción de reclamos es la misma para los más jóvenes y los más viejos.

Rp. a) H_0: reclamo indep de edad, b) H_0: las 4 proporciones de reclamos son iguales: $\chi_k^2 = 9.3314 = 3$, $R.C. =]7.82, +\infty[$, se rechaza H_0. significación=0.025 c) $p=0.15$, ES=0.0412, Z=(0.15−0.10)/ES=2.425, $RA=[-1.96, 1.96]$, son diferentes..

35. La fábrica de cerámicas "ARTE" realiza un estudio para determinar si la proporción de artículos defectuosos producidos por los trabajadores es la misma durante los turnos de mañana, tarde o noche. Una muestra de la producción de un día ha dado los siguientes resultados:

	Turnos		
	Mañana	Tarde	Noche
Defectuosos	41	20	57
No defectuosos	369	380	323

Al nivel de significación de 0.01,

a) ¿Se puede concluir que la proporción de artículos defectuosos producidos es la misma para los tres turnos?.

b) ¿Qué pares de proporciones poblacionales son diferentes?

Rp. H_0: las tres proporciones son iguales, $\chi_k^2 = 21.22$, gl. = 2, $R.C. =]9.21, +\infty[$, se rechaza H_0. (Significación: 0.000), b) Mañana con tarde y tarde con noche.

36*. Con los datos de la hoja de cálculo del *estudio socioeconómico de los estudiantes universitarios de Lima* (ver apéndice) construya una tabla de contingencia entre las variables: Lugar de origen y número de hermanos. ¿Hay alguna relación entre estas dos variables?.

Capítulo 12

ANALISIS DE VARIANZA

Introducción

El análisis de varianza es una técnica estadística para comprobar si son iguales las medias de más de dos poblaciones mediante el análisis y la comparación de diversos tipos varianzas muestrales insesgadas.

Esta técnica introducida por R. A. Fisher en 1920 se origina en aplicaciones agrícolas, de ahí que su lenguaje contenga términos de carácter agrícola tales como *bloques* o *parcelas* y *tratamientos* (que se utiliza como sinónimo de variables independientes). Posteriormente la técnica estadística del análisis de varianza ha encontrado aplicación en casi todas las disciplinas científicas y ha llegado a convertirse en un tema muy amplio.

El nombre a análisis de varianza (ANOVA por sus iniciales en inglés) que se da a esta prueba de varias medias, proviene del hecho de que este método se basa en la comparación de varianzas estimadas de diversas fuentes.

Cada método de análisis de varianza está asociado a un modelo matemático específico. Los modelos se clasifican según el número de variables que han de ser probadas. Si es una variable, el modelo se denomina de *clasificación simple o de un sólo factor*. Si son dos variables, el modelo se denomina de *clasificación doble o de dos factores*.

Los modelos aquí expuestos son los siguientes:

1) Modelo de clasificación simple o experimentos de un factor
   a) Completamente aleatorizado,
   b) aleatorizado por bloques

2) Modelo de clasificación doble o experimentos de dos factores:
   a) sin replicación
   b) con replicación

Los ejemplos y ejercicios de este texto se han resuelto utilizando el paquete estadístico didáctico *MCEST* creado por el autor.

Para un estudio más avanzado de las técnicas de análisis de varianza el interesado debe consultar publicaciones que tratan en forma más amplia de este importante tema.

12.1 Análisis de varianza de un factor: Diseño completamente aleatorizado.

Sea X una característica que se mide en k poblaciones (o tratamientos) diferentes. con medias respectivas $\mu_1, \mu_2, \ldots, \mu_k$ y varianzas respectivas: $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$.

Las suposiciones del ANOVA son:

1. Las k poblaciones son independientes (o las k muestras son independientes)
2. Cada una de las poblaciones tiene distribución normal, $N(\mu_i, \sigma_i^2)$
3. Las k varianzas son iguales a la varianza común σ^2. Esta condición de la homogeneidad de la varianza se comprueba mediante el contraste de Bartlett (ver por ejemplo referencia 16 página 497).

Las k poblaciones juntas constituyen una población mayor cuya media μ (media total o gran media) se define por:

$$\mu = \frac{\sum_{i=1}^{k} \mu_i}{k}.$$

Para cada $i = 1, 2, \ldots, k$, sea $X_{i1}, X_{i2}, \ldots, X_{in_i}$ una muestra aleatoria simple de tamaño n_i escogida en la *i-esima* población.

Estas k muestras constituyen los subgrupos que **se supone pues son independientes**. Es decir, con los supuestos hechos, cada una de estas k muestras consiste de n_i variables aleatorias independientes supuestas normales $N(\mu_i, \sigma^2)$. Por lo tanto, las variables aleatorias X_{ij} que denotan a la j-ésima observación de la *i-ésima* muestra ($i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$) son independientes y tienen cada una distribución normal $N(\mu_i, \sigma^2)$.

En el modelo de clasificación de un factor completamente aleatorizado los valores x_{ij} de las k muestras (j-ésima observación de la *i-ésima* muestra $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$) se registran en un arreglo tabular como el de la tabla 12.1,

Tabla 12.1. *Datos de k muestras aleatorias independientes*

	Tratamientos						
	1	2	...	i	...	k	
	x_{11} x_{12} $\vdots$ x_{1n_1}	x_{21} x_{22} $\vdots$ x_{2n_2}	... ...	x_{i1} x_{i2} $\vdots$ x_{in_2}	... ...	X_{k1} x_{k2} $\vdots$ x_{kn_k}	
Total	$T_{1\bullet}$	$T_{2\bullet}$	...	$T_{i\bullet}$	...	$T_{k\bullet}$	$T_{\bullet\bullet}$
n_i	n_1	n_2	...	n_i	...	n_k	n
Medias	$\bar{x}_{1\bullet}$	$\bar{x}_{2\bullet}$	...	$\bar{x}_{i\bullet}$	...	$\bar{x}_{k\bullet}$	$\bar{x}_{\bullet\bullet}$

en donde,

$T_{i\bullet}$ es la suma de datos de la muestra *i*.

$T_{\bullet\bullet}$ es el total de datos de las *k* muestras.

$n_1 + n_2 + \cdots + n_k = n$, es el total observado en las *k* muestras.

$\bar{x}_{i\bullet}$ es la media de la muestra *i*, (estimación insesgada de la media μ_i).

$\bar{x}_{\bullet\bullet}$ media total muestral (estimación insesgada de la media μ).

El modelo del ANOVA

Cada observación X_{ij} ($i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$), de la muestra se expresa en la forma:

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

en donde ε_{ij} mide la desviación del dato observado x_{ij} con respecto a la media μ_i. Esta desviación se denomina también ***error* o *residuo***. Dado que las variables aleatorias X_{ij} son independientes y tienen cada una distribución normal $N(\mu_i, \sigma^2)$, las ε_{ij} son, entonces, variables aleatorias independientes y tienen distribución normal $N(0, \sigma^2)$.

Por otro lado, cada media μ_i se desvía de la media total μ una cantidad $\alpha_i = \mu_i - \mu$. Este desvío es denominado, ***efecto* del *i*-ésimo tratamiento**. Observe que:

$$\sum_{i=1}^{k} \alpha_i = 0.$$

En resumen, el modelo de clasificación simple o de un factor completamente aleatorizado, es la ecuación:

$$X_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

en donde, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$, $\sum n_i = n$,

Las variables aleatorias X_{ij} son independientes y normales $N(\mu_i, \sigma^2)$.

Las variables aleatorias ε_{ij} son independientes y normales $N(0, \sigma^2)$.

μ es media total, y $\alpha_i = \mu_i - \mu$ es el efecto del tratamiento i.

Las hipótesis del ANOVA

La hipótesis nula H_0 consiste en afirmar que **las medias de las k poblaciones (o tratamientos) son iguales**, (o las k muestras provienen de la misma población). Esto, es:

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k$$

Dado que $\mu_i = \mu$ es equivalente a $\alpha_i = 0$, para $i = 1, 2, \ldots, k$, la hipótesis nula consiste también en afirmar que no hay efecto en todos los tratamientos, esto es:

$$H_0: \ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0.$$

La hipótesis alternativa es,

Para la primera forma, H_1 : No todas las medias son iguales.

Para la segunda forma, H_1 : Al menos una de las α_i no es igual a cero.

El análisis

La prueba de la hipótesis H_0 contra H_1 se basa en dos estimaciones independientes de la varianza poblacional común σ^2. Estas estimaciones se obtienen particionando la *suma de cuadrados total* (SCT) : $\sum\sum(x_{ij} - \bar{x}_{\bullet\bullet})^2$ en dos componentes. En efecto, de la relación:

$$x_{ij} - \bar{x}_{\bullet\bullet} = x_{ij} - \bar{x}_{i\bullet} + \bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}$$

se obtiene la siguiente **identidad de suma de cuadrados**

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$$

Es conveniente simbolizar esta partición de suma de cuadrados por:

$$SCT \quad = \quad SCE \quad + \quad SCC$$

Donde,

SCE es la suma de cuadrados del error (o dentro de los tratamientos).

SCC es la suma de cuadrados de las columnas (o entre los tratamientos).

Se verifican las siguientes esperanzas de sumas de cuadrados. (Ver por ejemplo, referencia 15 página 432):

$$E(SCE) = (n-k)\sigma^2$$

$$E(SCC) = (k-1)\sigma^2 + \sum_{i=1}^{k} n_i \alpha_i^2$$

$$E(SCT) = (n-1)\sigma^2 + \sum_{i=1}^{k} n_i \alpha_i^2$$

En consecuencia,

1) De la primera esperanza, resulta que $\dfrac{SCE}{n-k}$ es una estimación insesgada de la varianza σ^2, independientemente de que la hipótesis nula sea verdadera o falsa.

2) Si la hipótesis nula, H_0: $\alpha_i = 0$, $(i = 1, 2, \ldots, k)$, se supone verdadera, entonces, de la segunda y tercera esperanza resultan respectivamente que

   $\dfrac{SCC}{k-1}$ es una estimación insesgada de σ^2, y

   $\dfrac{SCT}{n-1}$ es una estimación insesgada de σ^2

Las tres estimaciones insesgadas de la varianza común σ^2, se denominan **cuadrados medios** y son denotados respectivamente para el error, para las columnas y para el total, por: *CME*, *CMC*, y *CMT*.

Es de esperar, entonces, que el cociente *CMC/CME* sea cercano a uno si la hipótesis nula es verdadera. Pero, si la hipótesis nula no es verdadera *CME* no cambia, mientras, que *CMC* será mayor. Esto implica que el cociente será mayor que la unidad. Si invertimos el razonamiento, se concluye que si *CMC/CME* es significativamente grande se puede concluir que las medias de las poblaciones son distintas.

Además:

La variable aleatoria, SCC/σ^2 se distribuye como chi-cuadrado con $k-1$ grados de libertad (χ^2_{k-1}).

La variable aleatoria, SCE/σ^2 se distribuye como chi-cuadrado con $n-k$ grados de libertad (χ^2_{n-k}).

En consecuencia, si la hipótesis nula es verdadera, el cociente;

$$F = \frac{SCC/[\sigma^2(k-1)]}{SCE/[\sigma^2(n-k)]} = \frac{CMC}{CME} \text{ se distribuye según } F(k-1, n-k).$$

Esto es, la variable aleatoria F tiene distribución F con $k-1$ y $n-k$ grados de libertad.

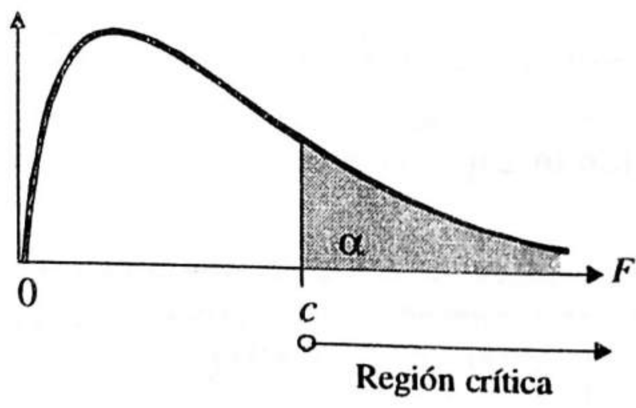
Además, dado el nivel de significación α, para los grados de libertad $k-1$ y $n-k$, en la tabla F se encuentra el *valor crítico* $c = F_{1-\alpha, k-1, n-k}$.

La **región crítica** o de rechazo de H_0 de la prueba es el intervalo $]c, +\infty[$ (ver figura 12.1)

A partir de los datos observados de la muestra se calcula: $F_{cal} = \dfrac{CMC}{CME}$.

La **regla de decisión** es: Rechazar la hipótesis nula H_0 si $F_{cal} > c$. En caso contrario no rechazar H_0. En el modo P, si $P = P[F > F_{cal}]$, se rechazará la hipótesis nula H_0 si $P < \alpha$. En caso contrario no se debe rechazar H_0.

Figura 12.1



Las sumas de cuadrados del total, de las columnas y del error se calculan utilizando las siguientes equivalencias:

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2 - C, \text{ donde, } C = \frac{T_{\bullet\bullet}^2}{n}$$

$$SCC = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 = \sum_{i=1}^{k}n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\frac{T_{i\bullet}^2}{n_i} - C,$$

$$SCE = SCT - SCC$$

Es práctico resumir las sumas de cuadrados, los grados de libertad, los cuadrados medios y la F calculada en la tabla 12.2 denominada **de análisis de varianza (ANOVA)**.

Tabla 12.2 ANOVA. *De un factor completamente aleatorizado.*

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados medios	Razón F calculada
Tratamientos (columnas)	SCC	$k-1$	$CMC=\dfrac{SCC}{k-1}$	$F_{cal}=\dfrac{CMC}{CME}$
Error	SCE	$n-k$	$CME=\dfrac{SCE}{n-k}$	
Total	SCT	$n-1$		

EJEMPLO 12.1 (Muestras de tamaños iguales).

El gerente de compras de la empresa "MODA" desea comparar la velocidad de 4 máquinas de marcas diferentes con el fin de adquirir la más veloz para su uso en una confección específica. Para esto, observó los tiempos que cada máquina utiliza para producir 6 unidades de la confección en forma aleatoria. Los tiempos registrados en segundos se presentan en la tabla 12.3.

Con un nivel de significación de 0.05, ¿es posible concluir que las máquinas utilizan la misma velocidad media por unidad de confección?.

Tabla 12.3. *Tiempos (en segundos) empleados en la producción.*

	Máquina				
	1	2	3	4	
	55	60	64	42	
	46	58	62	45	
	45	68	51	52	
	73	58	57	44	
	50	63	65	42	
	63	52	68	56	
Totales $T_{i\bullet}$	332	359	367	281	$T_{\bullet\bullet}=1339$
$n_i=r$	6	6	6	6	$n=24$
Medias $\bar{x}_{i\bullet}$	55.33	59.83	61.17	46.83	$\bar{x}_{\bullet\bullet}=55.79$

SOLUCION.

Sea μ_i la velocidad media de la máquina i, donde, $i=1, 2, 3, 4$.

1. *Hipótesis*: $H_0: \mu_1=\mu_2=\mu_3=\mu_4$

   H_1: No todas las medias son iguales

2. *Nivel de significación*: $\alpha=0.05$

3. *Estadística* $F = CMC/CME$ que se distribuye según $F(k-1, n-k)$, donde $k=4$ y $n=24$.
4. *Región crítica*: Para $\alpha = 0.05$, en la tabla F se encuentra el valor crítico de la prueba, $F_{0.95,3,20} = 3.10$. Se rechazará H_0, si $F_{cal} > 3.10$
5. *Cálculos*: De los datos se obtiene:

$$C = \frac{T_{\bullet\bullet}^2}{n} = \frac{(1339)^2}{24} = 74{,}705.04$$

$$SCT = \sum\sum x_{ij}^2 - C = (55)^2 + (46)^2 + \cdots + (56)^2 - 74{,}705.04 = 1851.96$$

$$SCC = \sum \frac{T_{i\bullet}^2}{r} - C = \frac{(332)^2 + (359)^2 + (367)^2 + (281)^2}{6} - 74{,}705.04 = 754.13$$

$$SCE = SCT - SCC = 1851.96 - 754.13 = 1{,}097.83\,.$$

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y la F calculada se dan en la tabla 12.4

Tabla 12.4 ANOVA. *Para los datos del ejemplo* 12.1

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados medios	Razón F Calculada
Máquinas (columnas)	754.125	3	251.375	$F = 4.579$
Error	1,097.833	20	54.892	
Total	1,851.958	23		

6. *Decisión*: Dado que $F = 4.579 > 3.10$, se rechaza, H_0.

NOTA. Con el paquete estadístico *MCEST* se obtiene la probabilidad P,

$$P = P[F > 4.579] = 0.013\,.$$

Dado que 0.013 < 0.05, se debería rechazar H_0.

EJEMPLO 12.2. (Muestras de tamaños diferentes)

Cuatro profesores P_1, P_2, P_3, y P_4, enseñan a muchos alumnos de su horario el mismo curso de Estadística. De uno de sus exámenes se extrajeron al azar una muestra de calificaciones de cada horario. Estas se registran de la siguiente manera:

Profesores			
P_1	P_2	P_3	P_4
12	14	13	10
11	16	12	17
09	13	09	15
10	18	11	14
	17	12	15
	12		

Al nivel de significación de 0.05, ¿se puede concluir que existe una diferencia significativa en las calificaciones promedio obtenidas con los cuatro profesores?.

SOLUCION.

Sea μ_i la media de las calificaciones del grupo i, $i = 1, 2, 3, 4$.

1. *Hipótesis*
   $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ contra H_1: No todas las medias μ_i son iguales
2. *Nivel de significación* : $\alpha = 0.05$.
3. *Estadística* $F = CMC/CME$ que se distribuye según $F(k-1, n-k)$, donde $k = 4$ y $n = 20$
4. *Región crítica*: Para el nivel $\alpha = 0.05$ en la tabla F se encuentra el valor crítico de la prueba es, $F_{0.95,3,16} = 3.24$. Se rechazará H_0, si $F_{cal} > 3.25$.
5. *Cálculos*: De los datos se obtienen:

$$T_{1\bullet} = 42,\ T_{2\bullet} = 90,\ T_{3\bullet} = 57,\ T_{4\bullet} = 71,\ T_{\bullet\bullet} = 260.$$

$$n_1 = 4,\ n_2 = 6,\ n_3 = 5,\ n_4 = 5,\ n = 20.$$

$$C = \frac{T_{\bullet\bullet}^2}{n} = \frac{(260)^2}{20} = 3{,}380$$

$$SCT = \sum\sum x_{ij}^2 - C = (12)^2 + (11)^2 + \cdots + (14)^2 + (15)^2 - 3{,}380 = 138$$

$$SCC = \sum \frac{T_{i\bullet}^2}{n_i} - C = \frac{(42)^2}{4} + \frac{(90)^2}{6} + \frac{(57)^2}{5} + \frac{(71)^2}{5} - 3{,}380 = 69$$

$$SCE = SCT - SCC = 138 - 69 = 69$$

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y la F calculada se resumen en la tabla 12.5

Tabla 12.5. ANOVA: *Para los datos del ejemplo* 12.2

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Razón *F* Calculada
Profesores	69.000	3	23.000	$F = 5.333$
Error	69.000	16	4.313	
Total	138.000	19		

6. *Decisión*: Dado que $F = 5.333 > 3.24$, se rechaza H_0.
   La probabilidad *P* de la prueba es, $P[F > 5.333] = 0.01$.

12.1.1 Comparación múltiple a posterior

Si se rechaza la hipótesis nula, $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, es decir, si la prueba *F* del análisis de varianza resulta significativa e indica que no todas las medias son iguales, es deseable determinar cuáles son las medias que no son iguales entre si.

La comparación de pares de medias se resuelve utilizando intervalos de confianza de pares de medias o ejecutando pruebas de hipótesis de pares de medias, procedimiento conocido como comparación múltiple a posteriori.

Observe que si k es el número de medias, se deben realizar con ellas $C_2^k = k(k-1)/2$ comparaciones de pares de medias.

El problema de las comparaciones múltiples ha sido objeto de muchas discusiones. Winer (1962) resumió varios (6) métodos para resolver este problema, algunos de ellos son más rigurosos que otros reduciendo la probabilidad de cometer error tipo I.

Uno de los métodos es la **diferencia mínima significativa (DMS)** que utiliza pruebas *t*-Student para llevar a cabo todas las comparaciones por pares entre las medias de los grupos. El defecto de este procedimiento es que si cada prueba *t* se realiza con probabilidad de error tipo I igual a 0.05, entonces, la probabilidad total de cometer un error tipo I será mayor que 0.05. Uno de los procedimientos para comparar medias conservando el nivel α original es la prueba *F* **del método de Scheffé**.

Cualquiera sea el método que se utilice para probar pares de medias, las hipótesis nula y alternativa a probar son:

$$H_0 : \mu_i = \mu_j \text{ contra } H_1 : \mu_i \neq \mu_j$$

para todo $i \neq j$, en donde $i, j = 1, 2, \ldots, k$

Si se utiliza la prueba *t*-Student, la estadística es la misma para probar igualdad de dos medias poblacionales normales con varianzas desconocidas que se suponen son iguales, es decir la estadística es:

$$t = \frac{\bar{x}_{i\bullet} - \bar{x}_{j\bullet}}{\sqrt{CME\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}}$$

cuya distribución es *t*-student con $n-k$ grados de libertad (los grados de libertad de *SCE*).

Entonces, se tienen los siguientes métodos para determinar pares de medias significativas:

1) **El intervalo de confianza** del $(1-\alpha)100\%$ para la diferencia de medias $\mu_i - \mu_j$ es:

$$(\bar{x}_i - \bar{x}_j) \mp t_0 \sqrt{CME(\frac{1}{n_i} + \frac{1}{n_j})}$$

Si el intervalo contiene el cero, se concluye que $\mu_i = \mu_j$.

2) **Prueba t de dos medias con varianzas iguales**: Si para el nivel de significación α en la tabla *t-S*tudent se encuentra el valor crítico, $c = t_{1-\alpha/2,n-k}$, entonces, se rechazará H_0 si el valor calculado de t es mayor que *c*. No se rechazará H_0 en caso contrario.

3) **Prueba DMS**. Dado que, $t^2_{1-\alpha/2,n-k} = F_{1-\alpha,1,n-k}$, (ver referencia 15 página 437), entonces, la variable

$$F = t^2 = \frac{(\bar{x}_{i\bullet} - \bar{x}_{j\bullet})^2}{CME} \frac{n_i n_j}{n_i + n_j}$$

se distribuye según una $F(1, n-k)$.

Si para el nivel de significación α en la tabla *F* se encuentra el valor crítico, $F_{1-\alpha,1,n-k}$, entonces, se rechazará H_0 si

$$\frac{(\bar{x}_{i\bullet} - \bar{x}_{j\bullet})^2}{CME} \frac{n_i n_j}{n_i + n_j} > F_{1-\alpha,1,n-k}$$

o si $$\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right| > \sqrt{\frac{n_i + n_j}{n_i n_j} \times CME \times F_{1-\alpha,1,n-k}}$$

no se rechazará H_0 en caso contrario.

Observe que, si los tamaños de muestras son todos iguales, digamos a un número r, esto es, $n_1 = n_2 =,\ldots,= n_k = r$, entonces, se rechazará H_0 si

$$\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right| > \sqrt{\frac{2}{r}(CME)F_{1-\alpha,1,\ n-k}}$$

EJEMPLO 12.3.

En el ejemplo 12.1 la decisión fue rechazar la hipótesis nula de que las 4 medias de grupos son iguales. Luego, existen diferencias significativas entre las marcas de las máquinas en cuanto a la velocidad promedio por unidad de confección. Al nivel de significación del 5% y utilizando la estadística F, ¿qué pares de medias son significativamente diferentes?.

SOLUCION.

El valor crítico para las comparaciones de todos los pares de medias de los grupos del método DMS es:

$$\sqrt{\frac{2}{r}(CME)F_{0.95,1,20}} = \sqrt{\frac{2}{6}(54.892)(4.35)} = 8.92$$

Se tiene entonces,

$\left|\bar{x}_{1\bullet} - \bar{x}_{2\bullet}\right| = |55.33 - 59.83| = 4.5 < 8.92$, no significativa.

$\left|\bar{x}_{1\bullet} - \bar{x}_{3\bullet}\right| = |55.33 - 61.17| = 5.84 < 8.92$, no significativa.

$\left|\bar{x}_{1\bullet} - \bar{x}_{4\bullet}\right| = |55.33 - 46.83| = 8.5 < 8.92$, no significativa.

$\left|\bar{x}_{2\bullet} - \bar{x}_{3\bullet}\right| = |59.83 - 61.17| = 1.34 < 8.92$, no significativa.

$\left|\bar{x}_{2\bullet} - \bar{x}_{4\bullet}\right| = |59.83 - 46.83| = 13 > 8.92$, * significativa.

$\left|\bar{x}_{3\bullet} - \bar{x}_{4\bullet}\right| = |61.17 - 46.83| = 14.34 > 8.92$, * significativa.

Si las muestras tienen el mismo tamaño las diferencias de medias, $\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right|$ para todo $i \neq j$ se pueden resumir como en la tabla 12.6, donde las filas y columnas son las medias de los grupos.

Tabla 12.6. *Comparaciones de pares de medias*

Grupos/ Medias		1 55.33	2 59.83	3 61.17	4 46.83
1	55.33	–			
2	59.83	4.5	–		
3	61.17	5.84	1.34	–	
4	46.83	8.5	13*	14.34*	–

Se indican con * las diferencias significativas al 5%. En este caso se tiene que 2 de los 6 pares de medias poblacionales son significativamente diferentes: $\mu_2 \neq \mu_4$, y $\mu_3 \neq \mu_4$.

Es evidente que se debe preferir la máquina 4 por ser la más veloz, es decir en promedio utiliza menos tiempo.

Si las muestras no tienen el mismo tamaño no existe un valor crítico único para realizar el contraste de pares de medias. En este caso se deben evaluar todos los posibles pares de medias en forma individual por la prueba t-Student.

El lector debería verificar que utilizando el método DMS para los datos del ejemplo 11.2 se tienen los siguientes pares de medias significativamente diferentes al 5%: $\mu_1 \neq \mu_2$, $\mu_1 \neq \mu_4$, $\mu_2 \neq \mu_3$ y $\mu_3 \neq \mu_4$.

NOTA. (Método a posteriori de Scheffé). El contraste posterior por el método de Scheffé, nos lleva a concluir que $\mu_i \neq \mu_j$ al nivel de significación α si,

$$F = \frac{(\bar{x}_{i\bullet} - \bar{x}_{j\bullet})^2}{(k-1) \times CME \times (1/n_i + 1/n_j)} > F_{1-\alpha,\, k-1,\, n-k}$$

o si

$$\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right| > \sqrt{(k-1) \times CME \times (1/n_i + 1/n_j) \times F_{1-\alpha,\, k-1,\, n-k}}$$

El valor crítico $F_{1-\alpha,\, k-1,\, n-k}$ es el mismo de la zona de rechazo en la prueba ANOVA. En el ejemplo 12.1 es $F_{0.95,\, 3,\, 20} = 3.10$.

Aplicando el método de Scheffé a los datos del ejemplo, 12.1 se encuentra que sólo la diferencia, $\bar{x}_{3\bullet} - \bar{x}_{4\bullet}$ es significativa, ya que:

$$F = \frac{(\bar{x}_{3\bullet} - \bar{x}_{4\bullet})^2}{(k-1) \times CME \times (1/n_3 + 1/n_4)} = \frac{(14.34)^2}{3 \times 54.892 \times (2/6)} = 3.75 > 3.10$$

Luego, al 5% se concluye por el método de Scheffé, que $\mu_4 \neq \mu_3$.

12.2 Análisis de varianza de un factor: Diseño aleatorizado por bloques

En el diseño completamente aleatorizado de un factor se ha supuesto que k tratamientos distintos se aplican a k muestras independientes. Uno de los tratamientos elegido al azar, era aplicado a una muestra aleatoria simple de tamaño n_1, otro de los tratamientos restantes elegido al azar a otra muestra aleatoria simple de tamaño n_2, etc., el último de todos los tratamientos se aplicaba a un muestra aleatoria simple de tamaño n_k.

En el análisis de varianza de un factor aleatorizados por bloques completos, todos los tratamientos son asignados aleatoriamente a los bloques. Un ejemplo típico para el diseño de bloques completamente aleatorizados, utilizando tres tratamientos T_1, T_2, T_3, asignados al azar a cuatro bloques, es como sigue:

Bloque 1	Bloque 2	Bloque 3	Bloque 4
T_3	T_2	T_1	T_3
T_2	T_1	T_3	T_1
T_1	T_3	T_2	T_2

Una vez terminado el experimento, los datos se registran en un arreglo 4×3 como se indica en el cuadro siguiente:

Bloques	Tratamientos		
	1	2	3
1	x_{11}	x_{21}	x_{31}
2	x_{12}	x_{22}	x_{32}
3	x_{13}	x_{23}	x_{33}
4	x_{14}	x_{24}	x_{34}

Donde x_{11} denota el resultado que se obtiene utilizando el tratamiento 1 en el bloque 1, x_{21} el que se obtiene utilizando el tratamiento 2 en el bloque 1, etc., x_{34} es el que se obtiene utilizando el tratamiento 3 en el bloque 4.

En general si k tratamientos A_i, $i = 1, 2, \ldots, k$ se asignan al azar a r bloques B_j, $j = 1, 2, \ldots, r$, los resultados se pueden resumir en un arreglo rectangular $r{\times}k$ como se muestra en la tabla 12.7 .

Tabla 12.7. *Datos muestrales del diseño por bloques aleatorios completos*

Bloques	Tratamientos						Total de Bloque	Medias de bloques
	A_1	A_2	$\ldots$	A_i	$\ldots$	A_k		
B_1	x_{11}	x_{21}	$\ldots$	x_{i1}	$\ldots$	x_{k1}	$T_{\bullet 1}$	$\bar{x}_{\bullet 1}$
B_2	x_{12}	x_{22}	$\ldots$	x_{i2}	$\ldots$	x_{k2}	$T_{\bullet 2}$	$\bar{x}_{\bullet 2}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
B_j	x_{1j}	x_{2j}	$\ldots$	x_{ij}	$\ldots$	x_{kj}	$T_{\bullet j}$	$\bar{x}_{\bullet j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
B_r	x_{1r}	x_{2r}	$\ldots$	x_{ir}	$\ldots$	x_{kr}	$T_{\bullet r}$	$\bar{x}_{\bullet r}$
Total	$T_{1\bullet}$	$T_{2\bullet}$	$\ldots$	$T_{i\bullet}$	$\ldots$	$T_{k\bullet}$	$T_{\bullet\bullet}$	
Medias	$\bar{x}_{1\bullet}$	$\bar{x}_{2\bullet}$	$\ldots$	$\bar{x}_{i\bullet}$	$\ldots$	$\bar{x}_{k\bullet}$		$\bar{x}_{\bullet\bullet}$

donde

$T_{i\bullet}$: Suma de datos de la *i*-ésima columna (tratamiento).

$T_{\bullet j}$: Suma de datos de la *j*-ésima fila (bloque).

$T_{\bullet\bullet}$: Suma de todas las *rk* observaciones.

$\bar{x}_{i\bullet}$: Media de los *r* datos observados en el tratamiento A_i.

$\bar{x}_{\bullet j}$: Media de los *k* datos observados en el bloque B_j.

$\bar{x}_{\bullet\bullet}$: Media de todas *rk* observaciones.

Cada una de las *rk* celdas contiene el dato x_{ij} que es el valor de una variable aleatoria X_{ij}.

Se supone que las variables aleatorias, X_{ij}, $i = 1, 2, \ldots, k$, y $j = 1, 2, \ldots, r$, son independientes y que cada una de ellas tiene distribución normal con media μ_{ij} y varianza común σ^2.

Observando la tabla 12.7, podemos concluir que en este caso se tiene una sola muestra de tamaño *r* a la que en un primer momento se le aplica un primer tratamiento elegido al azar, en un segundo momento le es aplicado un segundo tratamiento elegido al azar entre los restantes, etc., en un momento *k* le es aplicado el tratamiento restante *k*.

El modelo

Sea $\mu_{i\bullet}$ la media de las r medias poblacionales para el tratamiento i,

$$\mu_{i\bullet} = \frac{\sum_{j=1}^{r} \mu_{ij}}{r}, \quad i = 1, 2, \ldots, k$$

Análogamente, sea $\mu_{\bullet j}$ la media de las k medias poblacionales para el bloque j.

Esto es,

$$\mu_{\bullet j} = \frac{\sum_{i=1}^{k} \mu_{ij}}{k}, \quad j = 1, 2, \ldots, r$$

La media total μ de las rk medias poblaciones se define como ,

$$\mu = \frac{\sum_{i=1}^{k} \sum_{j=1}^{r} \mu_{ij}}{rk}$$

Cada observación X_{ij} puede escribirse en la forma

$$X_{ij} = \mu_{ij} + \varepsilon_{ij}$$

en donde ε_{ij} mide la desviación del dato observado X_{ij} de la media poblacional μ_{ij}. Como cada X_{ij} se supone normal $N(\mu_{ij}, \sigma^2)$, entonces, las variables aleatorias ε_{ij} se suponen independientes y normales $N(0, \sigma^2)$.

Las desviaciones de μ_{ij} con respecto a μ se deben tanto a efectos de tratamientos como de bloques. Se supone que se suman los efectos, entonces:

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

donde α_i es el **efecto del i-ésimo tratamiento** y β_j es el **efecto del j-ésimo bloque** y donde $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, r$. Se impone la restricción:

$$\sum_{i=1}^{k} \alpha_i = 0, \quad \text{y} \quad \sum_{i=1}^{r} \beta_i = 0$$

Resumiendo, el **modelo** del análisis de varianza de un factor aleatorizado por bloques es la ecuación:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

en donde se supone que:

1) Los efectos de los tratamientos y de los bloques son aditivos, es decir no hay efecto conjunto entre α_i y β_j. **No hay interacción** entre tratamientos y bloques.
2) Las variables aleatorias ε_{ij} son independientes y normales $N(0,\sigma^2)$.

NOTA. En el análisis de varianza de un factor aleatorizado por bloques, los bloques no deben ser considerados como un factor, ya que estos son considerados sólo como material experimental cuyos efectos influyen en los efectos de los tratamientos.

Las hipótesis

Para determinar si hay diferencias significativas entre las medias poblacionales de los k tratamientos (columnas) las hipótesis nula H_0^C y alternativa H_1^C respectivamente son:

$$H_0^C : \mu_{1\bullet} = \mu_{2\bullet} = \cdots = \mu_{k\bullet}$$
$$H_1^C : \text{No todas las } \mu_{i\bullet} \text{ son iguales}$$

Para determinar si hay diferencias significativas entre las medias poblacionales de los r bloques (filas) las hipótesis nula H_0^F y alternativa H_1^F respectivamente son:

$$H_0^F : \mu_{\bullet 1} = \mu_{\bullet 2} = \cdots = \mu_{\bullet r}$$
$$H_1^F : \text{No todas las } \mu_{\bullet j} \text{ son iguales}$$

Por otra parte, si suponemos que $\sum_{i=1}^{k}\alpha_i = 0$, y $\sum_{j=1}^{r}\beta_j = 0$, entonces,

$$\mu_{i\bullet} = \frac{\sum_{j=1}^{r}\mu_{ij}}{r} = \frac{\sum_{j=1}^{r}(\mu + \alpha_i + \beta_j)}{r} = \mu + \alpha_i$$

$$\mu_{\bullet j} = \frac{\sum_{i=1}^{k}\mu_{ij}}{k} = \frac{\sum_{i=1}^{k}(\mu + \alpha_i + \beta_j)}{k} = \mu + \beta_j$$

Luego, las hipótesis respectivas en función de los efectos de tratamientos y de los bloques son las siguientes:

Para tratamientos (columnas)

$H_0^C: \ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.

H_1^C: Al menos uno de los α_i no es igual a cero.

Para bloques (filas)

$H_0^F: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0$.

H_1^F: Al menos uno de los β_j no es igual a cero

El análisis

De la identidad:

$$x_{ij} - \bar{x}_{\bullet\bullet} = (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet}) + (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) + (\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet}) \quad ,$$

se obtiene;

$$\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2$$

Si se utiliza la notación:

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{\bullet\bullet})^2 , \quad SCE = \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2$$

$$SCC = \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 , \quad SCF = \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2$$

se tiene entonces,

$$SCT = SCE + SCC + SCF$$

Además, se verifica (ver referencia 16 página 521), que:

$$E(SCE) = (r-1)(k-1)\sigma^2$$

$$E(SCC) = (k-1)\sigma^2 + r\sum_{i=1}^{k}\alpha_i^2$$

$$E(SCF) = (r-1)\sigma^2 + k\sum_{j=1}^{r}\beta_j^2$$

Luego, si H_0^C es verdadera, entonces, el cuadrado medio de columnas,

$$CMC = SCC/(k-1) \text{ es un estimación insesgada de } \sigma^2.$$

Análogamente, si H_0^F es verdadera, entonces, el cuadrado medio de filas,

$$CMF = SCC/(r-1) \text{ es un estimación insesgada de } \sigma^2..$$

Por otra parte, independientemente de los valores de verdad de los dos tipos de hipótesis nulas de tratamientos y bloques, el cuadrado medio de errores,

$$CME = \frac{SCE}{(k-1)(r-1)} \text{ es un estimación insesgada de } \sigma^2.$$

Para probar la hipótesis nula de que los efectos de los tratamientos son todos iguales a cero, se utiliza la estadística:

$$F^C = \frac{CMC}{CME} \text{ que tiene distribución } F(k-1,(k-1)(r-1))$$

En forma semejante, para probar la hipótesis nula de que los efectos de los bloques son todos iguales a cero, se utiliza la estadística:

$$F^F = \frac{CMC}{CME} \text{ que tiene distribución } F(r-1,(k-1)(r-1))$$

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y las F calculadas se resumen en la tabla 12.8.

Tabla 12.8. ANOVA. *Modelo de clasificación simple aleatorizado por bloques.*

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Razón F Calculada
tratamientos (columnas)	SCC	$k-1$	$CMC = \frac{SCC}{k-1}$	$F^C = \frac{CMC}{CME}$
Entre bloques (filas)	SCF	$r-1$	$CMF = \frac{SCF}{r-1}$	$F^F = \frac{CMF}{CME}$
Error	SCE	$(r-1)(k-1)$	$CME = \frac{SCE}{(r-1)(k-1)}$	
Total	SCT	$rk-1$		

Para calcular las sumas de cuadrados, se ejecutan las siguiente equivalencias:

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}x_{ij}^2 - C, \quad \text{donde } C = \frac{T_{\bullet\bullet}^2}{rk}$$

$$SCC = \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 = \frac{1}{r}\sum_{i=1}^{k}T_{i\bullet}^2 - C,$$

$$SCF = \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2 = \frac{1}{k}\sum_{j=1}^{r}T_{\bullet j}^2 - C\,,$$

$SCE = SCT - (SCC + SCF)$

Los grados de libertad de SCE se obtiene también por sustracción:

$$(k-1)(r-1) = (rk-1) - (k-1) - (r-1)$$

NOTA. En el modelo simple completamente aleatorizado por bloques sólo interesa determinar si los efectos de los tratamientos son nulos o no. **No es importante docimar si los efectos de los bloques son nulos o no, por que los bloques se consideran simplemente como material experimental**. Sin embargo esta segunda prueba es importante para el modelo de clasificación de dos variables sin repetición como se verá en la siguiente sección.

EJEMPLO 12.4.

Se realiza un estudio para comparar cinco variedades de arroz (A_1, A_2, A_3, A_4, A_5) en cuanto a su rendimiento. Se cuentan con cuatro lugares de siembra de igual tamaño y fertilidad en San Martín. Para evitar que pueda producirse algún efecto por los diferentes lugares de siembra, se hizo un diseño aleatorizado por bloques asignando cada variedad al azar a cada uno de cuatro lugares. Se registraron los siguientes rendimientos en cientos de kilogramos.

Lugar 1	Lugar 2	Lugar 3	Lugar 4
$A_1 = 12$	$A_3 = 7$	$A_2 = 14$	$A_5 = 7$
$A_5 = 13$	$A_1 = 7$	$A_4 = 12$	$A_3 = 6$
$A_4 = 16$	$A_2 = 12$	$A_5 = 8$	$A_1 = 6$
$A_3 = 11$	$A_4 = 12$	$A_1 = 8$	$A_2 = 10$

Utilice un nivel de significación del 0.05 probar la hipótesis de que no existe diferencia en los rendimientos de las cinco variedades de arroz.

SOLUCION.

Los datos de la muestra se resumen en el arreglo de la tabla 12.9.

1. *Hipótesis*:

   H_0: $\alpha_i = 0$, $i = 1, 2, 3, 4, 5$ (los efectos de las variedades de arroz son 0)

   H_1 : Al menos uno de las α_i no es cero

2. *Nivel de significación:* $\alpha = 0.05$.

3. *Estadística*: $F^C = \dfrac{CMC}{CME} \sim F(k-1,(r-1)\times(k-1))$, con $k=5$, $r=4$.

4. *Región crítica*: Para el nivel de significación $\alpha = 0.05$ y grados de libertad 4 y 12, en la tabla F se encuentra, $F_{0.95,4,12} = 3.26$. Se rechazará H_0 si el valor calculado de F es mayor que 3.26.

Tabla 12.9. *Rendimientos de cinco variedades de arroz.*

Lugares (bloques)	Variedades de arroz					Total de Bloque $T_{\bullet j}$	Media de Bloque $\bar{x}_{\bullet j}$
	A_1	A_2	A_3	A_4	A_5		
1	12	15	11	16	13	67	13.4
2	7	12	7	12	8	46	9.2
3	8	14	7	12	8	49	9.8
4	6	10	6	13	7	42	8.4
Total $T_{i\bullet}$	33	51	31	53	36	$T_{\bullet\bullet} = 204$	
Medias $\bar{x}_{i\bullet}$	8.25	12.75	7.75	13.25	9.0		$\bar{x}_{\bullet\bullet} = 10.2$

5. *Cálculos*: De los datos, se obtiene:

$$C = \frac{T_{\bullet\bullet}^2}{rk} = \frac{(204)^2}{4\times 5} = 2080.8.$$

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{r} x_{ij}^2 - C = ((12)^2 + (7)^2 + \cdots + (7)^2) - 2080.8 = 191.2.$$

$$SCC = \frac{1}{r}\sum_{i=1}^{k} T_{i\bullet}^2 - C = \frac{(33)^2 + (51)^2 + \cdots + (36)^2}{4} - 2080.8 = 108.2$$

$$SCF = \frac{1}{k}\sum_{j=1}^{r} T_{\bullet j}^2 - C = \frac{(67)^2 + (46)^2 + (49)^2 + (42)^2}{5} - 2080.8 = 73.2.$$ $SCE =$ $SCT - (SCC + SCF) = 191.2 - (108.2 + 73.2) = 9.8$.

Tabla 12.10. ANOVA. *Para los datos del ejemplo* 12.4.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Razón F Calculada
Variedades de Arroz	108.2	5 – 1=4	27.050	$F^C = 33.122$
Lugares	73.2	4 – 1=3	24.400	
Error	9.8	4×3 = 12	0.817	
Total	191.2	20 – 1=19		

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y las F calculadas se resumen en la tabla 12.10

6. *Decisión*: Dado que $F^C = \dfrac{27.05}{0.817} = 33.122 > 3.26$, se debe rechazar H_0.

EJEMPLO 12.5.

En el ejemplo 12.4 la decisión es rechazar la hipótesis nula. Luego, existen diferencias significativas entre promedios de rendimientos de las variedades de arroz. Determine qué par o pares de medias difieren significativamente. Utilice el nivel de significación, $\alpha = 0.05$.

SOLUCION.

Para este problema $n = r = 4$. Al nivel de significación $\alpha = 0.05$ el valor crítico de la prueba de pares de medias poblacionales es:

$$\sqrt{\frac{2}{r}(CME)F_{0.95,1,12}} = \sqrt{\frac{2}{4}(0.817)(4.54)} = 1.36$$

Las diferencias de medias, $|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}|$ para todo $i \neq j$ se presenta en la tabla 12.11, donde las filas y columnas son las medias de los grupos. Representando las diferencias significativas al 5% por *, se tiene que 6 de los 10 pares son significativamente diferentes.

Tabla 12.11. *Comparaciones de pares de medias*

Grupos/ Medias	1 8.25	2 12.75	3 7.75	4 13.25	5 9.00
1 8.25	–				
2 12.75	4.5*	–			
3 7.75	0.5	5.0*	–		
4 13.25	5.0*	0.5	5.5*	–	
5 9.00	0.75	3.75*	1.25	4.25*	–

12.3. Análisis de varianza de dos factores.

En el análisis de varianza con dos criterios de clasificación los datos de la muestra son clasificados por medio de un arreglo rectangular en el cual las columnas representan los niveles de un factor A y las filas, los niveles de una factor B. Cada combinación de fila y de columna definen una celda en el arreglo. Entonces se tienen kr celdas.

El análisis de varianza de dos factores se clasifica según el número de observaciones en las celdas. Si cada celda contiene una sola observación de la muestra, el modelo se denomina sin réplicas (o sin repetición). En cambio, si cada celda contiene dos o más observaciones de la muestra, el modelo se denomina con réplicas (o con repetición).

El modelo de clasificación de dos factores y sin réplica es similar al modelo de clasificación de un solo factor aleatorizado en bloques. En este caso los niveles de uno de los factores son los bloques.

En el modelo clasificación de dos factores, las dos variables son independientes, es decir, **no hay interacción entre los dos factores**. Sólo hay interacción si se toman observaciones múltiples en las diversas combinaciones de los dos factores.

En el modelo de clasificación de dos factores con replicas los tratamientos no son independientes. En este caso si **hay interacción de los dos factores**.

12.3.1 Análisis de varianza de dos factores: sin replicación

El modelo de clasificación de dos factores A y B sin replicación (o repetición) se caracteriza por que cada celda del arreglo rectangular contiene una sola observación. Es decir, cada observación x_{ij} es una muestra de tamaño 1 extraída de una población correspondiente a la combinación de tratamientos A_iB_j, en donde, $i=1,2,\dots,k$, $j=1,2,\dots,r$.

La tabla del análisis de varianza y el proceso de prueba para el modelo de clasificación de dos factores sin interacción **son exactamente los mismos** a los del modelo de clasificación de un factor aleatorizado por bloques. En vez de bloques se tiene el segundo factor. Las pruebas de tratamientos y de bloques son ahora importantes y se deben realizar simultáneamente.

EJEMPLO 12.6.

Los artículos fabricados por una compañía se producen por 4 operarios utilizando 5 máquinas diferentes. El fabricante quiere determinar si hay diferencias significativas entre las máquinas y entre los operarios. Se efectuó un experimento para determinar el número de artículos diarios producidos por cada operario utilizando cada una de las máquinas. Los resultados se dan en la tabla 12.12.

Utilizando un nivel de significación del 5%, pruebe si existe una diferencia significativa;

a) Entre las máquinas,
b) Entre los operarios.

Tabla 12.12. *Número de unidades producidas por día.*

Operarios	Máquinas					Total de filas $T_{\bullet j}$	Media de filas $\bar{x}_{\bullet j}$
	A_1	A_2	A_3	A_4	A_5		
B_1	15	18	14	19	16	82	16.4
B_2	10	15	10	15	11	61	12.2
B_3	11	17	10	15	12	65	13.0
B_4	9	13	9	16	10	57	11.4
Total $T_{i\bullet}$	45	63	43	65	49	$T_{\bullet\bullet} = 265$	
Medias $\bar{x}_{i\bullet}$	11.25	15.75	10.75	16.25	12.25		$\bar{x}_{\bullet\bullet} = 13.25$

SOLUCION.

1. *Hipótesis*:

a) Para las máquinas A_i (columnas)

$H_0^C : \alpha_i = 0$, $i = 1, 2, 3, 4, 5$ (los efectos de todas las máquinas son 0)

H_1^C : Al menos uno de las α_i no es cero

b) Para los operarios B_j (filas)

$H_0^F : \beta_j = 0$, $j = 1, 2, 3, 4$ (los efectos de todos los operarios son 0)

H_1^F : Al menos uno de las β_j no es cero

2. *Nivel de significación:* $\alpha = 0.05$.

3. *Estadísticas*: Para , $k = 5$, $r = 4$, se tienen:

a) Para columnas es : $F^C = \dfrac{CMC}{CME} \sim F(k-1, (r-1)\times(k-1))$

b) Para filas es $F^F = \dfrac{CMF}{CME} \sim F(r-1, (r-1)\times(k-1))$.

4. *Región crítica*: Para $\alpha = 0.05$ los valores críticos respectivos de columna y fila son: $F^{C}_{0.95,4,12} = 3.26$ y $F^{F}_{0.95,3,12} = 3.49$. Se rechazará la hipótesis nula respectiva si el valor calculado de F es mayor que el valor crítico respectivo.

5. *Cálculos*: De los datos se obtiene,: $C = \dfrac{T^2_{\bullet\bullet}}{rk} = \dfrac{(265)^2}{4\times 5} = 3511.25$.

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{r} x_{ij}^2 - C = ((15)^2 + (10)^2 + \cdots + (10)^2) - 3511.25 = 187.75.$$

$$SCC = \frac{1}{r}\sum_{i=1}^{k} T^2_{i\bullet} - C = \frac{(45)^2 + (63)^2 + \cdots + (49)^2}{4} - 3511.25 = 106.2$$

$$SCF = \frac{1}{k}\sum_{j=1}^{r} T^2_{\bullet j} - C = \frac{(82)^2 + (61)^2 + (65)^2 + (57)^2}{5} - 3511.25 = 72.55.$$ $SCE = SCT - (SCC + SCF) = 187.75 - (106.0 + 72.55) = 9.20$

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y las F calculadas se resumen en la tabla 12.13

Tabla 12.13. ANOVA. *Para los datos del ejemplo* 12.6.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Razón F Calculada
Máquinas	106.00	5 – 1=4	26.500	$F^C = 34.565$
Operarios	72.55	4 – 1=3	24.183	$F^F = 31.544$
Error	9.20	4×3 = 12	0.767	
Total	187.75	20 – 1=19		

6. *Decisión*: Dado que $F^C = 34.565 > 3.26$, se debe rechazar H_0^C (significación 0.045) y dado que $F^F = 31.544 > 3.49$, se debe rechazar H_0^F (significación 0.000).

12.3.2 Análisis de varianza de dos factores: con replicación.

El análisis de varianza con dos criterios de clasificación, con réplicación, conocido también como diseño completamente aleatorizado de dos factores, se basa en un arreglo rectangular de las observaciones, en el que las c columnas representan los niveles o tratamientos del factor A y las r filas los niveles o tratamientos del factor B. Cada combinación de tratamiento define una celda en la tabla. Se tiene entonces rc celdas. Cada celda contiene n ($n \geq 2$) observaciones (réplicas). Los rcn datos se muestran en la tabla 12.14,

Tabla 12.14. *Experimento de dos factores con n réplicas*

Factor B	Factor A				Total de Bloque
	1	2	...	c	
1	x_{111} x_{112} ⋮ x_{11n}	x_{211} x_{212} ⋮ x_{21n}	...	x_{c11} x_{c12} ⋮ x_{c1n}	
	$T_{11\bullet}$	$T_{21\bullet}$		$T_{c1\bullet}$	$T_{\bullet 1\bullet}$
2	x_{121} x_{122} ⋮ x_{12n}	x_{221} x_{222} ⋮ x_{22n}	...	x_{c21} x_{c22} ⋮ x_{c2n}	
	$T_{12\bullet}$	$T_{22\bullet}$		$T_{c2\bullet}$	$T_{\bullet 2\bullet}$
...	...	...	...	...	...
r	x_{1r1} x_{1r2} ⋮ x_{1rn}	x_{2r1} x_{2r2} ⋮ x_{2rn}	...	x_{cr1} x_{cr2} ⋮ x_{crn}	
	$T_{1r\bullet}$	$T_{2r\bullet}$		$T_{cr\bullet}$	$T_{\bullet r\bullet}$
Total	$T_{1\bullet\bullet}$	$T_{2\bullet\bullet}$		$T_{c\bullet\bullet}$	$T_{\bullet\bullet\bullet}$

en donde,

x_{ijk} es la k-ésima observación del i-ésimo nivel del factor A y del j-ésimo nivel del factor B,: $i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, r$, $k = 1, 2, \ldots, n$

$T_{ij\bullet}$: Suma de datos de la ij-ésima celda.

$T_{i\bullet\bullet}$: Suma de datos de la i-ésima columna

$T_{\bullet j\bullet}$: Suma de datos de la j-ésima fila

$T_{\bullet\bullet\bullet}$: Suma de todas las rcn observaciones.

$\bar{x}_{i\bullet\bullet}$: Media de datos de la i-ésima columna

$\bar{x}_{\bullet j\bullet}$: Media de datos de la j-ésima fila

$\bar{x}_{\bullet\bullet\bullet}$: Media de todas las rcn observaciones.

En este caso los factores o las variables A y B no son independientes por lo tanto es posible que exista *interacción* entre los dos factores. La interacción indica que los efectos de los niveles de una factor varía con los niveles del otro factor.

Con el análisis de varianza de dos factores con réplicas, se pueden probar 3 hipótesis nulas distintas, estas son: que no existen efectos por columna (o las medias por columna no difieren significativamente) , que no existen efectos por filas (o las medias por renglón no difieren significativamente), y que no existe interacción entre los dos factores (los dos factores son independientes.

El modelo

El análisis de varianza de dos factores y con réplicas se basa en el modelo matemático:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

en donde, $i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, r$, $k = 1, 2, \ldots, n$

μ es la media global (sin importar el tratamiento)

α_i es el efecto del i-ésimo tratamiento del factor A.

β_j es el efecto del j-ésimo tratamiento del factor B.

γ_{ij} es el efecto de interacción del i-ésimo tratamiento del factor A y del j-ésimo tratamiento del factor B,

ε_{ijk} es el error aleatorio asociado al proceso de muestreo.

Se supone que: las variables aleatorias ε_{ijk} son independientes $N(0,\sigma^2)$ y que:

$$\sum_{i=1}^{c}\alpha_i = 0,\ \sum_{j=1}^{r}\beta_i = 0,\ \sum_{i=1}^{c}\gamma_{ij} = 0\ \ \forall j \text{ , y } \sum_{j=1}^{r}\gamma_{ij} = 0\ \forall i.$$

Las hipótesis

En el análisis de varianza de dos factores con replicación interesa realizar tres pruebas de hipótesis. Estas son: para columnas, para filas y para interacción. Las hipótesis nulas y alternativas son las siguientes:

a) $H_0^C : \alpha_i = 0$, $i = 1, 2, \ldots, c$. (Los efectos de todos los tratamientos del factor A o de todas las columnas son nulos)

$H_1^C : \alpha_i \neq 0$ para algunas columnas.

b) $H_0^F : \beta_j = 0$, $j = 1, 2, \ldots, r$. (Los efectos de todos los tratamientos del factor B o de todas las filas son nulos)

$H_1^F : \beta_j \neq 0$ para algunas filas.

c) $H_0^I : \lambda_{ij} = 0$, $i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, r$ (los efectos conjuntos en todas las celdas son nulos, o no hay interacción entre filas y columnas)

$H_1^I : \lambda_{ij} \neq 0$ para algunas celdas.

El análisis.

Las estadísticas para realizar las tres pruebas de las hipótesis nulas dadas se obtienen de la partición de suma de cuadrados siguiente:

$$\sum_{i=1}^{c}\sum_{j=1}^{r}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{\bullet\bullet\bullet})^2 = rn\sum_{i=1}^{c}(\bar{x}_{i\bullet\bullet} - \bar{x}_{\bullet\bullet\bullet})^2 + cn\sum_{j=1}^{r}(\bar{x}_{\bullet j\bullet} - \bar{x}_{\bullet\bullet\bullet})^2$$
$$+ n\sum_{i=1}^{c}\sum_{j=1}^{r}(\bar{x}_{ij\bullet} - \bar{x}_{i\bullet\bullet} - \bar{x}_{\bullet j\bullet} + \bar{x}_{\bullet\bullet\bullet})^2$$
$$+ \sum_{i=1}^{c}\sum_{j=1}^{r}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij\bullet})^2$$

Esta partición la representamos simbólicamente por:

$$SCT = SCC + SCF + SCI + SCE$$

en donde, SCI es la suma de cuadrados debido a la interacción.

Los grados de libertad respectivos son:

$$rcn - 1 = (c-1) + (r-1) + (r-1)(c-1) + rc(n-1)$$

Por otro lado, se verifica la distribución de las siguientes estadísticas:

$$F^C = \frac{SCC/(c-1)}{SCE/(rc(n-1))} = \frac{CMC}{CME} \sim F(c-1, rc(n-1))$$

$$F^F = \frac{SCF/(r-1)}{SCE/(rc(n-1))} = \frac{CMF}{CME} \sim F(r-1, rc(n-1))$$

$$F^I = \frac{SCI/((r-1)(c-1))}{SCE/(rc(n-1))} = \frac{CMI}{CME} \sim F((r-1)(c-1), rc(n-1)).$$

La estadística $F^C = \dfrac{CMC}{CME}$ se utiliza para probar la hipótesis de columnas.

La estadística $F^F = \dfrac{CMF}{CME}$ se utiliza para probar la hipótesis de filas.

La estadística $F^I = \dfrac{CMI}{CME}$ se utiliza para probar la hipótesis de interacción.

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y las F calculadas se resumen en la tabla 12.15

Tabla 12.15, ANOVA. *Modelo de clasificación a dos factores con repetición*

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Razón F Calculada
Factor A (Columnas)	SCC	$c-1$	$CMC=\frac{SCC}{c-1}$	$F^C=\frac{CMC}{CME}$
Factor B (Filas)	SCF	$r-1$	$CMF=\frac{SCF}{r-1}$	$F^F=\frac{CMF}{CME}$
Interacción $A\times B$	SCI	$(r-1)(c-1)$	$CMI=\frac{SCI}{(r-1)(c-1)}$	$F^I=\frac{CMI}{CME}$
Error	SCE	$rc(n-1)$	$CME=\frac{SCE}{rc(n-1)}$	
Total	SCT	$rcn-1$		

Para calcular las sumas de cuadrados se ejecutan las siguientes equivalencias:

$$SCT=\sum_{i=1}^{c}\sum_{j=1}^{r}\sum_{k=1}^{n}(x_{ijk}-\bar{x}_{\bullet\bullet\bullet})^2=\sum_{i=1}^{c}\sum_{j=1}^{r}\sum_{k=1}^{n}x_{ijk}^2-C\text{, donde } C=\frac{T_{\bullet\bullet\bullet}^2}{rcn}$$

$$SCC=rn\sum_{i=1}^{c}(\bar{x}_{i\bullet\bullet}-\bar{x}_{\bullet\bullet\bullet})^2=\frac{1}{rn}\sum_{i=1}^{c}T_{i\bullet\bullet}^2-C$$

$$SCF=cn\sum_{j=1}^{r}(\bar{x}_{\bullet j\bullet}-\bar{x}_{\bullet\bullet\bullet})^2=\frac{1}{cn}\sum_{j=1}^{r}T_{\bullet j\bullet}^2-C$$

$$SCE=\sum_{i=1}^{c}\sum_{j=1}^{r}\sum_{k=1}^{n}(x_{ijk}-\bar{x}_{ij\bullet})^2=\sum_{i=1}^{c}\sum_{j=1}^{r}\sum_{k=1}^{n}x_{ijk}^2-\frac{1}{n}\sum_{j=1}^{r}T_{ij\bullet}^2$$

$$SCI=SCT-(SC+SCF+SCE)$$

EJEMPLO 12.7.

Se realiza una investigación para comparar cuatro métodos de dieta a fin de determinar su eficacia en términos del peso perdido en kilos. Con este fin, se diseñó un modelo de análisis de varianza de dos factores, considerando el método de dieta como el factor A con 4 niveles (A_i método i, $i = 1, 2, 3, 4$) y el peso inicial como el factor B con tres niveles (B_1=moderadamente pesado, B_2=pesado, B_3=muy pesado). Se eligieron al azar a dos personas de B_1 para A_1, dos para A_2, etc.. Después de un mes la pérdida de peso en kilogramos de las 24 personas que se sometieron a las dietas se registró en la tabla 12.16.

Tabla 12.16. *Pesos perdidos en kg. de 24 personas.*

Peso inicial	Tipos de dieta				Total
	A_1	A_2	A_3	A_4	$T_{\bullet j \bullet}$
B_1	8 7	6 5	7 7	5 6	
Total A_i	15	11	14	11	51
B_2	4 3	5 4	3 4	4 4	
Total A_i	7	9	7	8	31
B_3	7 6	6 7	5 6	7 6	
Total A_i	13	13	11	13	50
Total $T_{i \bullet \bullet}$	35	33	32	32	$T_{\bullet\bullet\bullet}$=132

Utilice un nivel de significación de 0.05 para probar las hipótesis:

a) El tipo de dieta no afecta a la pérdida de peso.

b) El peso inicial no afecta a la pérdida de peso.

c) No hay interacción entre los tipos de dieta y los niveles de peso inicial.

SOLUCION.

1. *Hipótesis*:

Hipótesis nulas

c) $H_0^C : \alpha_i = 0$, para todas las dietas: $i = 1, 2, 3, 4$

d) $H_0^F : \beta_j = 0$, para los pesos iniciales: $j = 1, 2, 3$

e) $H_0^I : \lambda_{ij} = 0$, para todas las celdas: $i = 1, 2, 3, 4$, $j = 1, 2, 3$

Hipótesis alternativas respectivas

a) H_1^C : $\alpha_i \neq 0$ para algunas columnas.

b) H_1^F : $\beta_j \neq 0$ para algunas filas.

c) H_1^I : $\lambda_{ij} \neq 0$ para algunas celdas.

2. *Nivel de significación:* $\alpha = 0.05$

3. *Estadísticas*: Para columnas, filas, e interacción respectivamente son:

$$F^C = \frac{CMC}{CME} \sim F(3,\ 12),\ F^F = \frac{CMF}{CME} \sim F(2,\ 12),\ F^I = \frac{CMI}{CME} \sim F(6,\ 12).$$

3. *Regiones críticas*. Los valores críticos para columnas, filas, e interacción son respectivamente: $F_{0.95,3,12} = 3.49$, $F_{0.95,2,12} = 3.89$, $F_{0.95,6,12} = 3.00$. Se rechazará la hipótesis nula respectiva si el valor calculado de F es mayor que el valor crítico respectivo.

4. *Cálculos*. De la tabla 12.16, se obtiene:

$$C = \frac{T_{\bullet\bullet\bullet}^2}{rcn} = \frac{(132)^2}{3\times4\times2} = 726$$

$$SCT = \sum_{i=1}^{4}\sum_{j=1}^{3}\sum_{k=1}^{2} x_{ijk}^2 - C = (8)^2 + (4)^2 + \cdots + (3)^2 - 600 = 772 - 726 = 46$$

$$SCC = \frac{1}{rn}\sum_{i=1}^{c} T_{i\bullet\bullet}^2 - C = \frac{1}{3\times2}[(35)^2 + (33)^2 + (32)^2 + (32)^2] - 726 = 1$$

$$SCF = \frac{1}{cn}\sum_{j=1}^{r} T_{\bullet j\bullet}^2 - C = \frac{1}{4\times2}[(51)^2 + (31)^2 + (50)^2] - 726 = 31.75$$

$$SCE = \sum_{i=1}^{4}\sum_{j=1}^{3}\sum_{k=1}^{2} x_{ijk}^2 - \frac{1}{n}\sum_{j=1}^{r} T_{ij\bullet}^2 = 772 - \frac{1}{2}[(15)^2 + (11)^2 + \cdots + (13)^2] = 5$$

$$SCI = SCT - (SC + SCF + SCE) = 46 - (1 + 31.75 + 5) = 8.25$$

Tabla 12.17. ANOVA. *Para los datos del ejemplo* 12.7

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Razón F calculada
Máquinas (A)	1.000	3	0.333	$F^C = 0.799$
Operarios (B)	31.750	2	15.875	$F^F = 38.100$
Interacción $A \times B$	8.250	6	1.375	$F^I = 3.300$
Error de muestreo	5.000	12	0.417	
Total	46.000	23		

5. *Decisión*: Dado que

$$F^C = \frac{0.333}{0.417} = 0.799 < 3.49, \text{ se acepta } H_0^C$$

$$F^F = \frac{15.875}{0.417} = 38.1 > 3.89, \text{ se rechaza } H_0^F$$

$$F^I = \frac{1.375}{0.417} = 3.3 > 3.00, \text{ se rechaza } H_0^I$$

EJERCICIOS

1. Una organización de consumidores selecciona al azar 5 secadoras de ropa de cada una de tres fábricas importantes para efectuar un estudio de comparación de tiempo promedio de secado. Se tabuló el tiempo (en minutos) requerido por cada máquina para secar un lote estándar de ropa. De los datos se ha obtenido la siguiente tabla de análisis de varianza:

Fuente de varianzas	S.C.	G.L.	C.M.	F
Tipos de Secadoras Error	377.733			
Total	594.933			

a) Establezca las hipótesis nula y alternativa

b) Determine la región crítica de la prueba, al nivel de significación 0.05

c) Complete la tabla ANOVA, ¿cuál es su decisión respecto a la hipótesis nula?.

Rp. H_0: las 3 medias de ventas son iguales, SCE=217.200, gl: 2, 12, 14, RC=]3.89, +∞[, F=10.435, se rechaza H_0, Significación =0.002 .

2. En 15 tiendas se colocaron tres tipos de publicidad. Se asignaron 5 de estas tiendas al azar a cada uno de los tipos distintos de publicidad con el propósito de estudiar el impacto de los carteles en las ventas. Al cabo de un mes, el número de ventas de cada una de las cinco tiendas asignadas a cada uno de los tipos de publicidad dieron los siguientes resultados:

Suma de los cuadrados de todas las ventas:

$$\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}^2 = 96,698$$

Suma de los totales de ventas para cada tipo de publicidad:

$$T_{1\bullet} = 400, \quad T_{2\bullet} = 425, \quad T_{3\bullet} = 375.$$

Al nivel de significación de 0.05, ¿proporcionan estos datos suficiente evidencia para inferir que los promedios de ventas son diferentes para los tres tipos de carteles?.

Rp. H_0: las 3 medias de ventas son iguales, SCC=250, SCE=448, SCT=698, gl: 2, 12, 14, F=3.35, gl: 2, 12, $P[F>3.35]$=0.07, se acepta H_0

3. Se efectúa un experimento para determinar el rendimiento de 4 variedades de papa. Se dispone de 20 parcelas de igual fertilidad que se dividieron en 4 grupos de 5 parcelas cada una. En cada grupo se sembró una variedad distinta de papa. Los rendimientos en kg. por parcela se dan en la tabla que sigue.

Variedades de papa			
V1	V2	V3	V4
55	52	53	52
53	58	55	50
60	50	57	51
52	60	51	49
53	52	54	53

Pruebe la hipótesis de que no existen diferencias significativas entre las producciones medias de las 4 variedades de papa al nivel de significación 0.05.

Rp. H_0: las 4 medias de rendimientos son iguales, SCC=42.6, SCE=146.4, SCT=189
F=1.55, gl: 3, 16, $P[F>1.55]$=0.24. Se acepta H_0.

4. Una compañía desea comparar cuatro tipos de neumáticos. Se asignó aleatoriamente los neumáticos a seis automóviles semejantes. La duración de los neumáticos en miles de kilómetros se da en la tabla que sigue.

Tipos de neumáticos			
N1	N2	N3	N4
55	63	48	59
53	67	50	68
50	55	59	57
60	62	50	66
55	70	47	71
65	75	61	73

Al nivel de significación del 5%.

a) ¿Se puede concluir que existe alguna diferencia en los rendimientos medios de los tipos de neumáticos?.

b) Si se rechaza la hipótesis nula, utilice la prueba t para probar si la duración media de los neumáticos tipo 1 es distinta a la duración media de los neumáticos de tipo 4?.

Rp. a) H_0: las 4 medias de resistencias son iguales, SCC=781.46, SCE=769.5, SCT=1550.96
F=6.77, gl: 3, 20, $P[F>6.77]$=0.002. Se rechaza H_0.
b) t dos medias varianzas iguales, t=−2.7299, gl=10, significación=0.021.

6. Un promotor inmobiliario está considerando invertir en un centro comercial a construirse en una capital del interior del país. Se evalúan 4 ciudades: Arequipa, Iquitos, Piura, Trujillo, en donde es muy importante el nivel de ingresos mensuales de las familias. Para resolver este problema se diseñó una prueba hipótesis de varias medias seleccionando una muestra aleatoria de ingresos familiares en cada una de las ciudades, obteniendo los siguientes ingresos en cientos de dólares:

Ingresos mensuales en decenas de $			
Arequipa	Iquitos	Piura	Trujillo
61	71	56	50
56	73	61	40
49	66	47	50
55	61	51	50
	46	58	50
		62	40
		65	

a) ¿Cuáles son los supuestos para realizar la prueba ANOVA?

b) Con un nivel de significación de 0.05, ¿puede el promotor concluir que hay diferencias en el ingreso medio?

c) ¿Qué pares de medias muestrales son significativamente diferentes?. ¿En qué ciudad debería construir el centro comercial?

d) Utilice una prueba *t*, para probar que el ingreso medio en arequipa es diferente al ingreso medio en Trujillo.

Rp. b) H_0: $\alpha_i=0$ $i=1,2,3,4$, $SCC = 1083.666$, $SCE = 993.668$, $SCT = 2077.334$ $F=7.270$, gl: 3, 20, $P[F>7.27]=0.002$. Se rechaza H_0, c) 1 con 4, 2 con 4, y 3 con 4 (DMS), d) $\bar{x}_1=38.5$, $\hat{s}_1=6.442$, $\bar{x}_4=25$, $\hat{s}_4=5.477$, $s_C^2=35.748$, ES=3.452, $t_K=3.91$, RC={$T<-2.228$ o $T>2.228$}, se rechaza H_0.

5. Pruebe la hipótesis de que las medias de las poblaciones 1, 2, 3, y 4 son iguales con un nivel de significación de 0.05, con base en las siguientes muestras aleatorias escogidas de las poblaciones respectivas.

Muestras			
Muestra 1	Muestra 2	Muestra 3	Muestra 4
15	10	20	10
20	12	08	08
08	20	20	16
14	13	15	15
18	18	10	
14	20	11	
	12	13	
	17	12	
	14		

Indique las hipótesis nula y alternativa, la regla de decisión, el cuadro de ANOVA y su decisión respecto a la hipótesis nula.

Rp. H_0: las 4 medias poblacionales son iguales. $SCC=27.727$, $SCE=378.347$, $SCT=406.074$ $F=0.56$, gl: 3, 23, $P[F>0.56]=0.65$. Se acepta H_0.

7. El Decano de la facultad desea estudiar el número de horas que los alumnos de los ciclos: 5, 6, 7 y 8, utilizan los terminales de computo de la universidad. Una muestra de usos por ciclo ha dado los siguientes tiempos en horas mensuales:

Ciclos			
C5	C6	C7	C8
35	43	28	39
33	47	30	48
30	35	39	37
40		30	46
35		27	
42			

Con un nivel de significación de 0.05.

a) ¿Es posible inferir que hay diferencias significativas en el número medio de horas de uso mensuales por ciclo de los terminales de computo?

b) Si hay diferencias significativas entre las medias de uso de los terminales, ¿qué pares de medias de los ciclos producen la diferencia?.

Rp. a) ANOVA: SCC=383.478, SCE=349.300, SCT=735.778, GL:3, 14, 17, CM: 128.826, 24.950, F=5.163, significación=0.013, b) Prueba DMS de pares de medias significativas, 2 con 3, 10.87, 3 con 2,−10.87, 4 con 3,11.7.

8. Se ha aplicado tres métodos diferentes de enseñanza a tres grupos de alumnos de estadística, el primero compuesto por 8 estudiantes, el segundo de 6 y el tercero de 12. Se quiere saber si estos métodos tienen algún efecto sobre las notas. Las calificaciones sobre un máximo de 20 obtenidas por dichos alumnos en una prueba final se dan en la tabla que sigue:

Métodos		
A	*B*	*C*
13	17	10
14	16	11
12	16	15
13	17	10
12	17	14
15	13	13
11		10
14		13
		11
		14
		13
		10

a) Con un nivel de significación de 0.05, ¿hay diferencia en las calificaciones promedio entre los tres métodos?.
b) Si se rechaza la hipótesis nula, realice pruebas adicionales por el método de intervalo de confianza para determinar que pares de calificaciones promedio son diferentes.

Rp. a) H_0: las 3 medias son iguales, $SCC = 64.62$, $SCE = 62$, $SCT = 126.62$, F=11.985, gl: 2, 23, $P[F>11.985]$=0.0003. Se rechaza H_0, b) 1 con 2 y 2 con 3.

9. Dieciséis personas fueron distribuidas aleatoriamente en 4 grupos distintos de 4 personas cada uno. A cada grupo se le asignó aleatoriamente un tiempo de entrenamiento antes de realizar cierta tarea. Los resultados de dicha tarea en los tiempos correspondientes se dan en la siguiente tabla.

Entrenamiento			
Grupo1 1 hora	Grupo 2 1.5 horas	Grupo 3 2 horas	Grupo 4 2.5 horas
25	14	7	8
19	26	10	7
22	17	9	9
20	15	11	4

a) Con estos datos y al nivel de significación 0.01, ¿se puede rechazar la hipótesis nula de que en la población los 4 grupos rindan igual? .
b) Si se rechaza la hipótesis nula, realice la prueba posterior DMS para determinar qué pares de medias son diferentes.

Rp. a) H_0: α_i=0 i=1,2,3,4, $SCC = 575.19$, $SCE = 133.75$, $SCT = 708.94$ $F = 17.2$, gl: 3, 12, $P[F>17.2]$=0.0001. Se rechaza H_0, b) 1 con 3, 1 con 4, 2 con 3 y 2 con 4 (DMS).

10. Veinte personas que experimentaban fiebres de 38 grados o más fueron divididos en 4 grupos de 6 personas cada uno y a cada grupo se le administró una marca de tableta distinta para el dolor de cabeza. El número de horas de descanso se da en la siguiente tabla:

Tableta			
T_1	T_2	T_3	T_4
5	11	6	12
3	5	4	10
8	3	7	9
4	3	5	8
2	4	6	7
6	2	2	8

Con los datos de esta muestra y utilizando el nivel de significación $\alpha = 0.05$.

a) Pruebe la hipótesis de que el número de horas promedio de descanso proporcionadas por las tabletas es el mismo para las 4 marcas.

b) Si se rechaza la hipótesis nula, ¿qué pares de medias muestrales son significativamente diferentes a ese nivel?.

Rp. a) H_0: las 4 medias son iguales, $SCC = 80.67$, $SCE = 108.67$, $SCT = 189.33$ F=4.95, gl: 3, 20, $P[F>4.95]$=0.01. Se rechaza H_0, b) 1 con 4, 2 con 4, y 3 con 4.

11. La estructura financiera de una firma se refiere a la forma en que se dividen los activos de la empresa por debe y haber, y el apalancamiento financiero se refiere al porcentaje de activos financiados por deuda. En un estudio financiero se afirma que el apalancamiento financiero puede utilizarse para aumentar la tasa de rendimiento sobre la inversión, es decir que, los accionistas pueden recibir rendimientos más altos con la misma cantidad de inversión gracias a su uso. Los siguientes datos muestran las tasas de rendimiento utilizando 3 diferentes niveles de apalancamiento financiero y un nivel de control (deuda cero) de 20 empresas seleccionadas al azar:

Tasas de rendimiento			
Control	Bajo	Medio	Alto
4.6	2.0	7.0	7.9
2.0	7.4	4.5	6.8
6.8	1.8	11.6	5.8
4.2	3.2	6.0	9.2
1.6	4.0	6.8	11.0

a) ¿Existen diferencias reales entre las medias de los cuatro niveles de rendimiento al nivel de significación 1%?, y al 5%?

b) ¿Son las tasas medias de rendimiento en los niveles de apalancamiento financiero bajo, medio y alto más altas que la del nivel de control?. Utilice DMS y Scheffé

Rp. a) H_0: α_i=0, i=1,2,3,4, SCC=65.588, SCE=84.920, SCT=150.508, F=4.119, g.l.=3,16, $P[F>4.119]$=0.024 al 1% se acepta H_0, al 5% se rechaza H_0 , b) Con DMS, 1 con 3, 1 con 4, 2 con 3 , 2 con 4, con Scheffé no hay pares significativos al 5%.

12. Se comparan tres máquinas A, B y C procedentes de diferentes fabricantes pero que producen el mismo tipo de objeto. Se quiere determinar si hay diferencias en la cantidad de objetos que producen por hora. Se tienen tres operarios. Para evitar cualquier sesgo por la diferente habilidad de los operarios se empleó un diseño aleatorizado por bloques estableciendo tres turnos para cada uno de los operarios asignándoles las tres máquinas en forma aleatoria. El número de objetos producidos se da en la siguiente tabla.

Operario 1	
B	18
A	23
C	17

Operario 2	
C	15
B	25
A	28

Operario 3	
A	16
C	16
B	22

Al nivel de significación 0.05, ¿es posible concluir que no existen diferencias entre las capacidades de producción de las tres variedades de papa?

Rp. H_0: $\alpha_i=0$ $i=1,2,3$, $SCC = 72.667$, $SCF=34.667$, $SCE=64.667$, $SCT=172$, $F=2.247$, gl: 2, 4, $P[F>2.247]=0.222$, se acepta H_0.

13. Se lleva a cabo un experimento en el cual se comparan tres tratamientos T_1, T_2, y T_3 en cuatro bloques. Se generó el siguiente diseño aleatorizado por bloques:

Bloque 1	Bloque 2	Bloque 3	Bloque 4
$T_2 = 50$	$T_3 = 63$	$T_1 = 46$	$T_3 = 64$
$T_1 = 45$	$T_2 = 52$	$T_3 = 62$	$T_1 = 44$
$T_3 = 60$	$T_1 = 45$	$T_2 = 53$	$T_2 = 52$

Al nivel de significación de 0.05,

a) Y Realizando el ANOVA de un factor aleatorizado por bloques, ¿podemos concluir que hay diferencias entre las medias de los tratamientos?
b) Si se rechaza la hipótesis nula, realice comparaciones apareadas de los promedios.
c) Continuando con b), ¿cuál de los tratamientos es el mejor?.

Rp. a) H_0: las 4 medias son iguales, $SCC = 604.5$, $SCF = 7.33$, $SCE = 8.17$, $SCT=620$, $F=222.061$, gl: 2, 6, $P[F>222.061]=0.000$, se rechaza H_0, b) 1 con 2, 1 con 3, 2 con 3, c) el mejor es T_3.

14. Se realiza un estudio de movimientos para determinar el mejor de tres métodos de montar un mecanismo. Para esto se diseñó un experimento de un factor por bloques aleatorios seleccionando cinco operarios con igual velocidad. El número de montajes terminados diarios por cada operario y con cada método se da en la tabla que sigue.

Operario	Método		
	1	2	3
1	3	9	5
2	4	8	6
3	3	7	8
4	5	9	7
5	4	6	9

Al nivel de significación del 5%

a) ¿Se puede concluir que los tres métodos de montaje son significativamente diferentes?.

b) Si se rechaza la hipótesis nula, ¿qué pares de métodos son significativamente diferentes a ese nivel?.

Rp. a) H_0: las 3 medias iguales, $SCC = 44.8$, $SCF = 3.067$, $SCE = 16.533$, $SCT=64.4$, $F=10.839$, gl: 2, 8, $P[F>10.839]=0.005$, se rechaza H_0, b) 1 con 2 y 1 con 3.

15. Los siguientes datos representan las calificaciones finales obtenidas por 4 alumnos de Estudios Generales Ciencias de la PUCP en Matemática I (M), Estadística (E), Física (F) y Lengua (L) :

Alumno	Cursos			
	M	*E*	*F*	*L*
1	14	13	14	16
2	13	18	15	19
3	11	16	13	18
4	12	15	12	16

Con un nivel de significación de 0.05

a) ¿Podemos concluir que los cursos presentan la misma dificultad?

b) Si se rechaza la hipótesis nula, ¿qué pares de cursos producen la diferencia?.

Rp. a) H_0: las 4 medias son iguales, $SCC=53.687$, $SCF=14.187$, $SCE=15.563$, $SCT=83.438$, $F = 10.349$, gl: 3, 9, $P[F>10.349]=0.003$, se rechaza H_0. b) 1 con 2, 1 con 4, y 3 con 4 (DMS).

16. Cuatro hombres realizan una tarea idéntica de embalaje de cajas. Se trata de determinar si hay una diferencia en la cantidad embalada de cajas debido a la habilidad de los hombres o debido al horario. Para esto se diseñó un análisis de varianza de dos factores. El número de cajas embaladas por cada hombre en 3 horas elegidas, se muestra en la tabla siguiente.

Horas	Hombres			
	H_1	H_2	H_3	H_4
10 – 11 a.m.	30	35	32	33
2 – 3 p.m.	18	21	20	19
4 – 5 p.m.	24	25	24	24

a) Plantee la hipótesis nula y alternativa para determinar si existen diferencias en el embalaje debido a los hombres, al nivel de significación de 0.01.

b) Plantee la hipótesis nula y alternativa para determinar si existen diferencias en el embalaje debido a los horarios, al nivel de significación de 0.01.

c) Si existen diferencias en alguno de los dos factores, ¿qué pares de niveles de éste, producen la diferencia?.

Rp. $SCC=13.583$, $SCF=346.167$, $SCE=5.167$, $SCT=364.917$, a) para las columnas, $F^C=5.258$, gl: 3, 6, $P[F>5.258]=0.04$, se acepta igualdad, b) para las filas, $F^F=201.$, gl: 2, 6, $P[F>201]=0.000$ se rechaza igualdad. c) Sólo en filas 1 con 2, 1 con 3, y 2 con 3.

17. Una empresa tiene 4 plantas en Sudamérica: Argentina, Brasil, Chile y Perú. Cada planta produce con 3 tipos diferentes de máquinas. Una muestra aleatoria del número de unidades producidas por planta y por máquinas se da en la siguiente tabla:

	Argentina	Brasil	Chile	Perú
Máquina 1	230	250	180	120
Máquina 2	160	180	120	80
Máquina 3	120	100	70	50

Aplicando un análisis de varianza de dos factores y el nivel de significación 0.05, determinar si existe alguna diferencia en la producción media

a) Debido a las máquinas,

b) Debido a los países.

Rp. SCC=17166.667, SCF=24266.667, SCE=1733.333, SCT=43166.667, gl: 3, 2, 6, 11,
a) columnas, $F^C = 19.808$, gl: 3, 6, $P[F>19.808]=0.002$, No se rechaza igualdad,
b) filas, $F^F = 42.000$, gl: 2, 6 , $P[F>42]=0.000$ se rechaza igualdad.

18. Una compañía que produce un tipo de artículo cuenta con 5 máquinas: M_1, M_2, M_3, M_4, y M_5 y con operarios: O_1, O_2, O_3 y O_4. Para comprobar si hay una diferencia en la cantidad de producción debido a la clase de máquinas y a la clase de operarios se diseñó un experimento de dos factores sin replicación asignando a cada operario una máquina por día. El número de artículos producidos se da en la tabla que sigue.

Operarios	Máquinas				
	M_1	M_2	M_3	M_4	M_5
O_1	23	25	30	32	40
O_2	28	27	35	38	42
O_3	32	30	37	39	43
O_4	36	38	40	43	45

Al nivel de significación $\alpha = 0.05$.

a) ¿Se puede concluir que existen diferencias en la producción debido a las máquinas?

b) ¿Se puede concluir que existen diferencias significativas en la producción debido a los operarios?

c) Si existen diferencias entre máquinas, ¿qué pares de máquinas producen la diferencia?.

Rp. SCC=471.8, SCF=282.55, SCE=36.20, SCT=790.55, g.l.:4, 3, 12, 19,
a) $H_0^C : \alpha_i = 0$ columnas, F^C=39.1, gl: 4, 12, $P[F>39.1]=0.000$, se rechaza H_0^C,
b) $H_0^F : \beta_j = 0$ filas, F^F=31.22, gl: 3, 12 , $P[F>39.1]=0.000$ se rechaza H_0^F.
c) 1 con 4, 1 con 5, 2 con 4, 2 con 5, y 3 con 5 (DMS).

19. Se diseñó un experimento de dos factores para probar la duración de focos de luz de 4 marcas diferentes fabricados con 4 tipos diferentes de filamentos. Los tiempos de duración para cada combinación de marcas y filamentos se dan en la tabla que sigue:

	Marcas			
Filamentos	A_1	A_2	A_3	A_4
B_1	390	402	392	385
B_2	380	403	394	385
B_3	377	411	399	380
B_4	370	404	400	384

Al nivel de significación 0.05

a) ¿Existe alguna diferencia en la duración media debido a las marcas?.
b) ¿Existe alguna diferencia en la duración media debido a los filamentos?.
c) Determine las diferencias para pares de medias de duración debido a las marcas

Rp. SCC=1671.5, SCF=18.5, SCE=300, SCT=1990, g.l.: 3, 3, 9, 15,

a) $H_0^C : \alpha_i = 0$ columnas, F^C=16.715, gl: 3,9, $P[F>16.715]$=0.001, se rechaza H_0^C.

b) $H_0^F : \beta_j = 0$ filas, F^F=0.185, gl: 3, 9 , $P[F>0.185]$=0.904, se acepta H_0^F.

c) 1 con 2, 1 con 3, 2 con 3, 2 con 4, y 3 con 4.

20. Los datos sobre ventas en miles de dólares con y sin publicidad, y con y sin descuento para un producto de consumo popular en ocho ciudades asignadas al azar se registraron para el diseño completamente aleatorizado de 2 factores en la tabla que sigue.

	Con publicidad	Sin publicidad
Con descuento	110.5 90.7	80.5 60.8
Sin descuento	50.4 40.6	35.8 20.7

a) Plantee las hipótesis respectivas
b) Al nivel de significación del 5% pruebe si existe efecto de los dos factores y de la interacción entre ellos.

Rp. SCC=1113.920, SCF=4753.125, SCI=80.645, SCE=552.090, SCT=6499.780, g.l.=1, 1, 1, 4, 7.

a) $H_0^C : \alpha_i = 0$ columnas, F^C=8.071, g.l.=1, 4, $P[F>8.071]$=0.047, se rechaza H_0^C.

b) $H_0^F : \beta_j = 0$ filas, F^F=34.437 g.l.=1, 4 $P[F>34.437]$=0.004, se rechaza H_0^F.

c) $H_0^I : \lambda_{ij} = 0$ interacción, F^I=0.584, g.l.=26, $P[F>0.584]$=0.487, se acepta H_0^I.

21. Se diseñó un experimento de dos factores con mediciones repetidas en terrenos de igual fertilidad para probar las diferencias entre 3 clases de maíz y 2 tipos de fertilizantes. De la producción se tomaron muestras aleatorias de tamaño $n = 2$ para cada combinación de maíz y fertilizante obteniendo los resultados de la tabla siguiente.

Fertilizante	Maíz		
	A_1	A_2	A_3
B_1	35 30	38 35	33 32
B_2	31 32	40 43	36 40

Al nivel de significación $\alpha = 0.05$, ¿proporcionan los datos suficiente evidencia para concluir que hay

a) Diferencias entre las medias de los tres tipos de maíz?.
b) Diferencias entre las medias de los tres tipos de fertilizantes?.
c) Interacción entre los tipos de maíz y tipos de fertilizantes?.

Rp. SCC=98.167, SCF=30.083, SCI=26.167, SCE=30.5, SCT=184.917, g.l.=2,1,2,6,11, a) $H_0^C : \alpha_i = 0$ columnas, F^C=9.656, g.l.=2,6, $P[F>9.656]$=0.013, se rechaza H_0^C.

b) $H_0^F : \beta_j = 0$ filas, F^F=5.918 g.l.=1,6 $P[F>5.918]$=0.051, se acepta H_0^F.

c) $H_0^I : \lambda_{ij} = 0$ interacción, F^I=2.574, g.l.=26, $P[F>2.574]$=0.156, se acepta H_0^I.

22. Se ha probado la vida útil de 3 marcas de pilas A_1, A_2, A_3 clasificados según sus costos B_1 (0.6\$), B_2(0.8\$), B_3(1.0\$). Los datos (en centenas de horas) se dan en la siguiente tabla.

Costos	Marcas de pilas		
	A_1	A_2	A_3
B_1	1.1 1.0	1.5 1.6	0.9 0.8
B_2	0.9 0.7	1.2 1.1	0.8 0.7
B_3	1.3 1.4	1.3 1.2	1.0 0,9

Al nivel de significación del 5% pruebe el efecto de los dos factores y de la interacción entre ellos.

Rp. SCC = 0.654, SCF = 0.288, SCI = 0.229, SCE = 0.060, SCT = 1.231, gl=2, 2, 4, 9, 17,

a) $H_0^C : \alpha_i = 0$ columnas, F^C = 49.083, gl=2, 9, $P[F>49.083]$=0.000, se rechaza H_0^C.

b) $H_0^F : \beta_j = 0$ filas, F^F =21.583, gl=2, 9, $P[F>21.583]$=0.000, se rechaza H_0^F.

c) $H_0^I : \lambda_{ij} = 0$ Interacción, F^I=8.583, gl=4, 9, $P[F>8.583]$=0.004, se rechaza H_0^I.

23. Para producir cierto bien una firma dispone de 4 máquinas de marcas distintas (A_i) que producen con igual velocidad y de 3 fuentes distintas de materia prima (B_j) de igual calidad. No se sabe si el número de unidades defectuosas producidas es la misma para las máquinas y para las materias primas. Se hizo operar cada marca de máquina con cada tipo de material durante 2 horas y se registró el siguiente número de unidades defectuosas por hora.

Materia prima	Máquinas A_1	A_2	A_3	A_4
B_1	6 5	4 3	5 5	3 4
B_2	2 1	3 2	1 2	2 2
B_3	5 4	3 4	3 4	4 3

Al nivel de significación de 0.05:

a) ¿Podemos concluir que hay diferencias significativas entre las máquinas A_i?

b) ¿Es posible concluir que hay diferencias significativas entre las materias primas B_j?

c) ¿Se puede inferir que hay efecto de interacción $A \times B$?

Rp. *SCC*=2.333, *SCF*=27.083, *SCI*=6.917, *SCE*=5.000, *SCT*=41.333, g.l.=3, 2, 6, 12, 23,

a) $H_0^C : \alpha_i = 0$ columnas, F^C=1.867, g.l.=3,12, $P[F>1.867]$=0.189, se acepta H_0^C.

b) $H_0^F : \beta_j = 0$ filas, F^F=32.5, g.l.=2, 12 $P[F>32.5]$=0.000, se rechaza H_0^F.

c) $H_0^I : \lambda_{ij} = 0$ interacción, F^I =2.767, g.l.=6, 12, $P[F>2.767]$=0.063, se acepta H_0^I.

24. Deseamos comprobar si la región geográfica y los ingresos familiares influyen en las puntuaciones obtenidas en una prueba nacional de inteligencia. Para esto, se eligieron de cada una de las 4 regiones 3 personas con ingresos bajos, 3 con ingresos medios y 3 con ingresos altos. Los resultados obtenidos por las 36 personas en la prueba de inteligencia vienen dados en la tabla que sigue.

Región geográfica	Ingresos Alto	Medio	Bajo
Norte	10 16 14	16 18 16	16 17 16
Sur	15 12 11	15 17 18	15 14 15
Centro	16 13 19	11 16 13	14 16 15
Oriente	17 18 17	16 18 18	13 14 13

Utilice el nivel de significación α=0.01 para probar las hipótesis

a) Es nulo el efecto debido a los ingresos familiares

b) Es nulo el efecto debido a las regiones .

c) Es nulo el efecto debido a la interacción entre región e ingresos.

Rp. SCC=10.889, SCF=10.444, SCI=76.222, SCE=72.667, SCT=170.222, g.l.=2, 3, 6, 24, 35.

a) $H_0^C : \alpha_i = 0$ **columnas,** F^C**=1.798, g.l.=2, 24,** $P[F>1.798]$**=0.187, se acepta** H_0^C **.**

b) $H_0^F : \beta_j = 0$ **filas** F^F**=1.15 g.l.=3, 24,** $P[F>1.15]$**=0.349, se acepta** H_0^F **.**

c) $H_0^I : \lambda_{ij} = 0$ **interacción,** F^I**=4.196, g.l.=6, 24,** $P[F>4.196]$**=0.005, se acepta** H_0^I **.**

25. Con los datos de la hoja de cálculo del *estudio socioeconómico de los estudiantes universitarios de Lima* (ver apéndice)

a) Y con un nivel de significación de 0.01, ¿existe diferencia en la edad media entre los tres grupos de estudiantes que provienen de la costa sierra o selva?.

b) Crea usted una variable que clasifique en tres grupos los gastos mensuales en educación: Menos de 1000, de 1000 a menos 3000, de 3000 o más. Con un nivel de significación de 0.05, ¿existe alguna diferencia entre el gasto medio de los tres grupos?

c) Crea usted una variable que clasifique en tres grupos los salarios mensuales: Menos de 2000, de 2000 a menos 4000, de 4000 o más. Con un nivel de significación de 0.05, ¿existe alguna diferencia entre el salario medio de los tres grupos?

Capítulo 13

REGRESION LINEAL Y CORRELACION

Introducción.

En muchas aplicaciones estadísticas se deben resolver problemas que contienen un conjunto de variables y que se sabe existe alguna asociación entre ellas. En este conjunto de variables muy a menudo se tiene una sola ***variable dependiente*** (o ***respuesta***) Y, que depende de una o más **variables independientes** o predictoras (o ***de regresión***) $X_1, X_2, \ldots, X_k$, como por ejemplo, el salario, dependiente de: años de experiencia, grado de instrucción, sexo, etc.

La variable dependiente se mide con un error que no se controla en el experimento, por tanto, **Y es una variable aleatoria**. Las variables independientes $X_1, X_2, \ldots, X_k$ se miden con un error despreciable, que en la mayoría de los casos se controla en el experimento, y por lo tanto, no tienen la propiedad de ser variables aleatorias.

Existen dos formas distintas pero relacionadas del estudio de la asociación entre variables a partir de una muestra aleatoria.

La primera forma, es *determinar una relación funcional de la variable dependiente Y con respecto a una o más variables independientes con el fin de predecir valores de Y.* Este método es el **análisis de regresión.**

La segunda forma de estudio de la asociación entre variables, es, *medir la magnitud relación entre ellas, mediante un coeficiente o índice.* A esta técnica se denomina **análisis de correlación**.

Los métodos de regresión y correlación entre variables se clasifican por el número de variables independientes, en simple y múltiple. El análisis de asociación se denomina **simple,** si hay una sola variable independiente, si hay dos o más variables independientes se denomina el **análisis de asociación múltiple**.

Por el tipo de función matemática que se puede ajustar a los datos, la asociación de las variables puede ser **lineal** o **no lineal** como por ejemplo, parábola, polinomio, exponencial, hipérbola etc. .

Los ejemplos y ejercicios de este capítulo han sido resueltos utilizando el paquete de computo estadístico didáctico *MCEST* creado por el autor de este libro.

Para un estudio más avanzado del análisis estadístico multivariado el lector debe consultar publicaciones que tratan en forma más amplia de este importante tema.

13.1 Regresión lineal simple.

13.1.1 Modelo de regresión lineal simple

Consideremos una variable dependiente Y con una sola variable independiente X. Representemos una muestra aleatoria de tamaño n de (X,Y) por el conjunto de pares de datos: $\{(x,y)/i=1,\ 2,\ldots,n\}$.

Si se toman muestras aleatorias adicionales utilizando exactamente los mismos valores de X, es de esperar que los valores de Y varíen. Por lo tanto, el valor y_i del par ordenado (x_i, y_i) será un valor de alguna variable aleatoria Y_i.

Es decir, para cada valor de X, hay un grupo de valores de Y

Por conveniencia, denotaremos por Y/X la variable aleatoria Y dependiente de X. Su media y varianza se denotan respectivamente por $\mu_{Y/X}$ y por $\sigma^2{}_{Y/X}$.

En particular el símbolo Y/x_i representa a la variable aleatoria Y_i cuando $X=x_i$. La media y varianza de Y/x_i son respectivamente μ_{Y/x_i} y $\sigma^2{}_{Y/x_i}$.

Los supuestos para el modelo regresión lineal simple son:

1) **Igualdad de varianzas (*Homoscedasticidad*).**

Para cada valor x_i de la variable independiente X, la distribución de la variable aleatoria dependiente Y_i tiene media μ_{Y/x_i} y varianza $\sigma^2{}_{Y/x_i}$. Se supone que cada una de estas varianzas son iguales a la varianza común σ^2, denominada **varianza de la regresión.** Es decir, las distribuciones de Y_i tienen medias diferentes, pero tienen la misma varianza σ^2.

2) **Independencia**

Se supone que las Y_i son variables aleatorias estadísticamente independientes.

3) **Linealidad.**

Se supone que la relación de Y con X es lineal, es decir todas las medias μ_{Y/x_i} deben estar en una línea recta denominada **línea de regresión poblacional**, cuya ecuación es, (ver figura 131.0)

$$\mu_{Y/X} = \alpha + \beta X$$

En la ecuación de regresión poblacional los *coeficientes de regresión* α y β son parámetros que se estiman a partir de los datos de la muestra..

El valor de α es la *ordenada en el origen* e indica el valor de Y cuando X = 0.

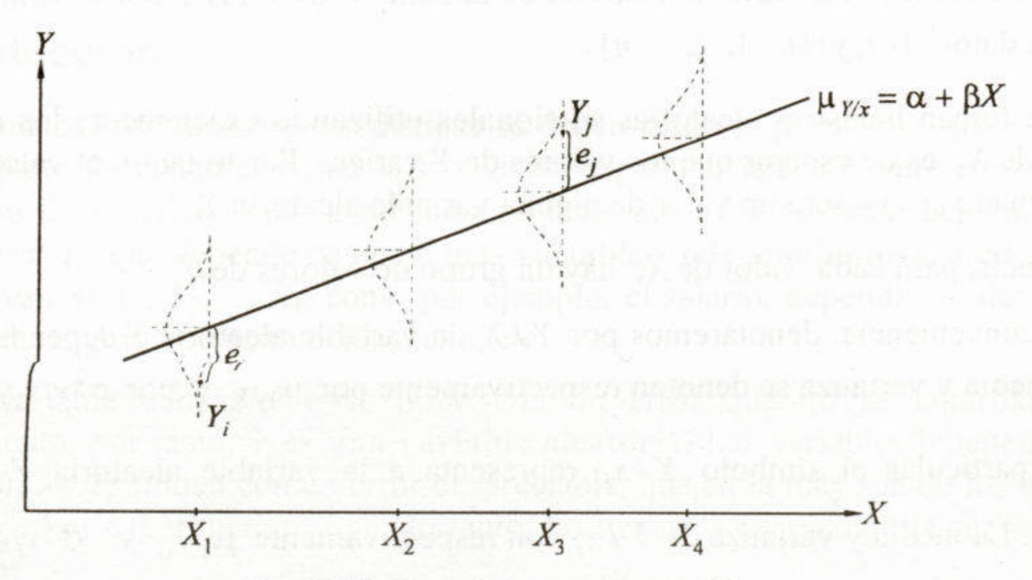


Fig. 131.0. *Suposiciones en regresión.*

El valor de β es la *pendiente de la ecuación de regresión* poblacional e **indica el cambio promedio en Y correspondiente a un incremento unitario en X.** El signo de β también indica el tipo de tendencia (positiva o negativa) de Y con respecto a X.

Cada valor individual Y_i difiere de la media condicional μ_{Y/x_i} en el término E_i, denominado ***error o residual***.

Las diferencias, $E_i = Y_i - \mu_{Y/x_i}$, son entonces, variables aleatorias independientes con media cero y varianza σ^2.

Luego, el modelo de regresión lineal simple puede ser expresado por la ecuación:

$$Y_i = \mu_{Y/x_i} + E_i = \alpha + \beta x_i + E_i$$

La estimación de la **ecuación de regresión poblacional** $\mu_{Y/X} = \alpha + \beta X$ es ***la ecuación de regresión muestral***:

$$\hat{Y} = a + bX$$

donde las **estadísticas** a y b son las estimaciones de los parámetros α y β, respectivamente.

Para inferencias acerca de β se requiere la suposición adicional de:

4) **Normalidad.**

Se supone que cada variable aleatoria dependiente Y_i tiene distribución normal con media μ_{Y/x_i} y varianza σ^2. En consecuencia, la distribución de cada variable E_i es normal con media 0 y varianza σ^2. (ver figura 131.0)

Diagrama de dispersión

El primer paso en el análisis de regresión, es construir una gráfica de los datos muestrales en el plano coordenado *XY*. Esta gráfica es denominada diagrama de dispersión.

El diagrama de dispersión indica frecuentemente el tipo de tendencia de *Y* con respecto a *X*.

Si la tendencia es lineal se puede ajustar una línea recta al diagrama de dispersión.

Por ejemplo, en las figuras 131.1 a), y b) los datos visualizan una *relación lineal* entre las variables *X* e *Y*. En la figura 131.1 c) los datos visualizan una relación, pero, una *relación no lineal o curvilinea*, y en la figura 131.1 d) los datos visualizan ninguna relación válida entre las variables.

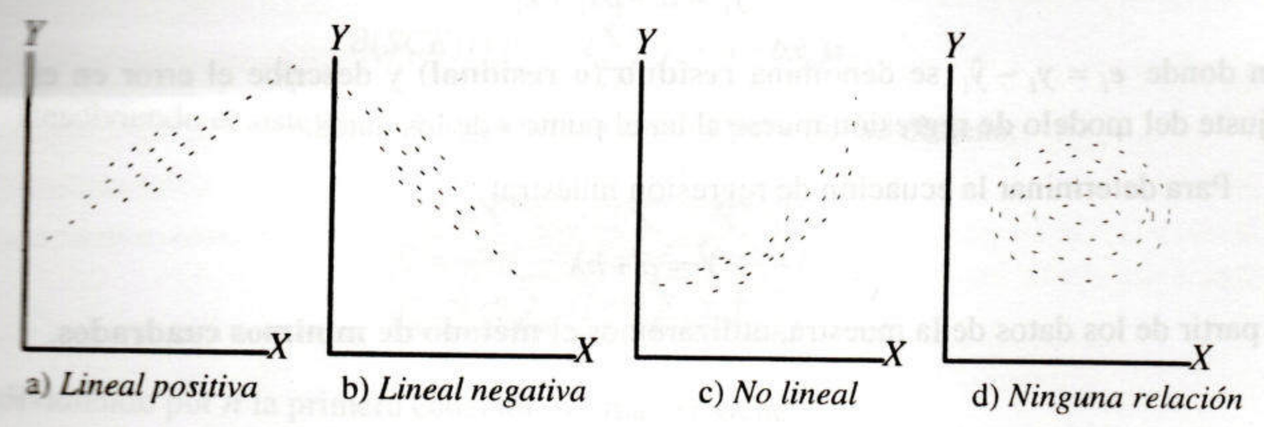


Fig. 131.1. *Diagramas de dispersión: Relaciones entre X e Y.*

13.1.2 Estimación de la ecuación de regresión poblacional. Método de mínimos cuadrados

En regresión lineal simple, una muestra de n pares de observaciones seleccionadas al azar de la población (*X*, *Y*) puede representarse por el conjunto: $\{(x, y)/i = 1,\ 2, \ldots, n\}$.

Si se trata de la **ecuación de regresión poblacional**,

$$\mu_{Y/X} = \alpha + \beta X\ .$$

cada dato (x_i, y_i) satisface la ecuación:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

en donde ε_i es el valor de E_i cuando Y_i toma el valor y_i.

Esta ecuación puede considerarse como **el modelo para una sola observación** y_i.

La estimación de la ecuación de regresión poblacional es **la ecuación de regresión muestral**:

$$\hat{Y} = a + bX$$

en donde $\hat{Y}$ es una estimación de $\mu_{Y/X}$, a y b son las estimaciones de los parámetros α y β, respectivamente.

Se denotará por $\hat{y}_i$ el valor de $\hat{Y}$ cuando $X = x_i$

Cuando se considera la ecuación de regresión muestral, cada dato (x_i, y_i) de la muestra, satisface la ecuación:

$$y_i = a + bx_i + e_i$$

en donde $e_i = y_i - \hat{y}_i$ se denomina **residuo (o residual)** y describe el error en el ajuste del modelo de regresión muestral en el punto i de los datos.

Para determinar la ecuación de regresión muestral

$$\hat{Y} = a + bX\ ,$$

a partir de los datos de la muestra, utilizaremos el **método de mínimos cuadrados**.

Método de mínimos cuadrados

La recta de regresión de mínimos cuadrados de Y en X es aquella que hace mínima la suma de los cuadrados de errores o residuos alrededor de la línea de regresión (SCE) cuya expresión es:

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Luego, determinar una recta de regresión de mínimos cuadrados consiste en hallar los valores de a y b de manera que hagan mínima, la suma:

$$SCE = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

Este requisito se cumple, de acuerdo con el teorema de Gass-Markow, si a y b se determinan resolviendo el siguiente sistema de ***ecuaciones normales***:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Estas ecuaciones **se obtienen de igualar a cero las derivadas de *SCE* con respecto a *a* y con respecto a *b*** respectivamente consideradas como variables, ya que (x_i, y_i) son datos observados.

En efecto, derivando con respecto a a y b, se tiene:

$$\partial(SCE)/\partial a = -2\sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\partial(SCE)/\partial b = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i$$

Resolviendo el sistema de ecuaciones normales para b, se obtiene:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

y dividiendo por *n* la primera ecuación normal, se tiene:

$$a = \overline{y} - b\overline{x}$$

Donde $\overline{x}$ e $\overline{y}$ son las medias de X e Y respectivamente.

Observe que: $b = \dfrac{\text{Covarianza de } X \text{ e } Y}{\text{Varianza de } Y} = \dfrac{s_{XY}}{s_X^2}$

NOTA 1. Sustituyendo $a = \overline{y} - b\overline{x}$ en: $Y = a + bX$, resulta,

$$Y - \overline{y} = b(X - \overline{x})$$

Esta es otra forma de expresar la recta de regresión. Observe que la recta de regresión contiene al punto $(\overline{x}, \overline{y})$ cuyas componentes son las medias de X y de Y respectivamente.

NOTA 2. (Interpretación del coeficiente de regresión *b*)

El valor constante a de la ecuación de regresión muestral, es la ordenada en el origen.

El valor de la pendiente b es el cambio promedio en $\hat{Y}$ cuando X cambia una unidad de medición

Si $b > 0$, entonces, la tendencia lineal es creciente, es decir, a mayores valores de X corresponden mayores valores de Y. También, a menores valores de X corresponden menores valores de Y.

Si $b < 0$, entonces, la tendencia lineal es decreciente, es decir, a mayores valores de X corresponden menores valores de Y. También, a menores valores de X corresponden mayores valores de Y.

Si $b = 0$, entonces, $Y = a$. Luego, Y permanece estacionario para cualquier valor de X. Es decir, **no hay regresión muestral**

Esta misma interpretación es válida para la pendiente β en la ecuación de regresión poblacional.

EJEMPLO 13.1.

El gerente de personal de la empresa agroindustrial 'Bajo Mayo" estudia la relación entre la variable dependiente: Y=gastos y la variable independiente X=salario, de su personal obrero. Una muestra aleatoria de 10 obreros reveló los siguientes datos en dólares por semana:

Salarios	28	25	35	40	45	50	50	35	70	80
Gastos	25	20	32	37	40	40	45	30	55	60

a) Trace el diagrama de dispersión e indicar la tendencia

b) Determine la línea de regresión muestral de mínimos cuadrados

c) De su comentario sobre el valor de la pendiente

d) Estime el gasto que correspondería a un salario semanal de $90.

SOLUCION.

a) El diagrama de dispersión de la muestra de ingresos (X) y gastos (Y) semanales es la figura 131.2.

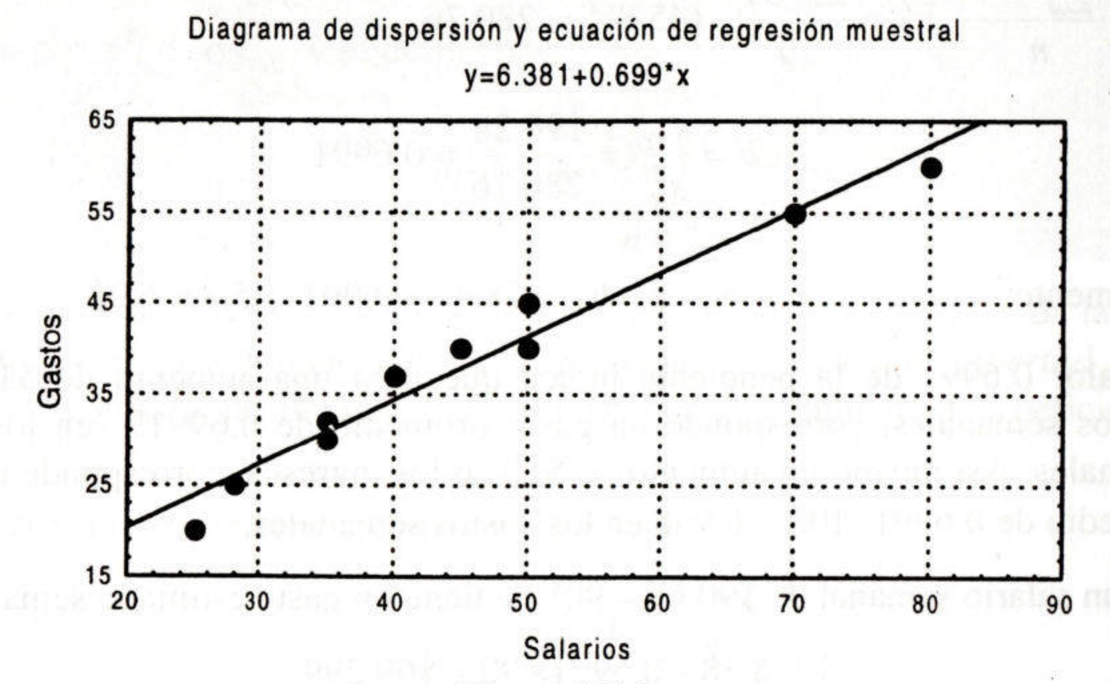


Figura 131.2

b) La línea de regresión estimada o ecuación de regresión muestral es:

$$\hat{Y} = 6.38 + 0.6991X$$

Esta ecuación puede ser obtenida usando una calculadora o un paquete de computo (por ejemplo, el *MCEST*).

El cálculo en forma detallada es como sigue:

$$\sum x = 458, \qquad \sum y = 384,$$

$$\sum xy = 19{,}550, \qquad \sum x^2 = 23{,}784$$

$$\bar{x} = 45.8, \qquad \bar{y} = 38.4$$

La pendiente b se puede obtener usando la formula de sumas

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10(19550) - (458)(384)}{10(23784) - (458)^2} = 0.6991$$

O usando la formula de covarianza y varianza

$$s_{XY} = \frac{\sum XY}{n} - \overline{xy} = \frac{19550}{10} - 45.8 \times 38.4 = 196.28$$

$$s_X^2 = \frac{\sum X^2}{n} - \bar{x}^2 = \frac{23784}{10} - (45.8)^2 = 280.76$$

$$b = \frac{s_{XY}}{s_X^2} = \frac{196.28}{280.76} = 0.6991$$

Finalmente: $a = \bar{y} - b\bar{x} = 38.4 - 0.6991 \times 45.8 = 6.38$

c) El valor 0.6991 de la pendiente indica que para una aumento de \$1 en los salarios semanales, corresponde un gasto promedio de 0.6991\$ en los gastos semanales. Así mismo un aumento de \$10 en los ingresos, corresponde un gasto promedio de $0.6991 \times 10\$ = 6.99\$$ en los gastos semanales.

d) Para un salario semanal de \$90 ($X = 90$), se tiene un gasto estimado semanal de

$$\hat{y} = 6.38 + 0.6991 \times 90 = \$69.299.$$

13.1.3 Estimación de la varianza de la regresión poblacional σ^2.

Una vez hallada la línea recta de regresión muestral, $\hat{Y} = a + bX$ nos interesa saber su utilidad. La utilidad principal es predecir valores de *Y* para valores determinados de *X*.

Si se hace una predicción nos interesa saber, ¿qué tan buena o confiable es esa predicción?. La respuesta a esta pregunta depende de la variabilidad de los valores de *Y* con respecto a la recta de regresión.

Una medida que indica el grado de variabilidad o dispersión (o concentración) en torno a la línea de regresión **es la varianza de la regresión poblacional**, que se denota por σ^2 o por $\sigma^2_{Y/X}$ y se define por:

$$\sigma^2 = E(Y - \mu_{Y/X}) = \frac{\sum_{i=1}^{N} (y_i - \mu_{Y/X})^2}{N}$$

donde N es el tamaño de la población.

La raíz cuadrada σ de esta varianza es la *desviación estándar de la regresión en la población*.

Una *estimación insesgada* de σ^2 es la **varianza de la regresión muestral** que se denota por s^2 o $\hat{\sigma}^2_{Y/X}$ y se define por:

$$s^2 = \frac{SCE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

donde, el numerador; es la suma de cuadrados de los errores alrededor de la línea de regresión y el denominador, $n-2$, representa los *grados de libertad* (a n se sustraen dos grados de libertad que corresponden al número de coeficientes de regresión).

Se verifica (ver referencia 16 página 386) que:

$$E(s^2) = \frac{E(SCE)}{n-2} = \sigma^2$$

Es decir que la varianza muestral s^2 es una estimación insesgada de σ^2.

Para el cálculo de s^2 se utiliza la siguiente expresión:

$$s^2 = \frac{SCE}{n-2} = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-2}$$

Error estándar de estimación

La raíz cuadrada de la varianza de la regresión muestral, es la *desviación estándar muestral de la regresión* denotada por s o por $\hat{\sigma}_{Y/X}$. Este valor se denomina también **error estándar de estimación.**

Interpretación del error de estimación

El error estándar de la estimación (o la varianza) es una medida de la dispersión de los valores observados alrededor de la ecuación de regresión muestral.

Mientras más pequeño sea el valor del error estándar de estimación s (o de la varianza s^2), más cercanos a la línea de regresión estarán los valores estimados $\hat{Y}$

Si la dispersión alrededor de la línea de regresión tiene distribución normal y el tamaño de la muestra es grande, entonces:

i) Si se traza una línea paralela a la línea de regresión $\hat{\sigma}_{Y/X}$ unidades más arriba y otra paralela $\hat{\sigma}_{Y/X}$ unidades más abajo, entonces, aproximadamente el 68% de los puntos del diagrama de dispersión caerán entre los valores $\pm\hat{\sigma}_{Y/X}$. Es decir, aproximadamente el 68% de los residuos son menores de $\pm\hat{\sigma}_{Y/X}$

ii) Si se traza una línea paralela a la línea de regresión $2\hat{\sigma}_{Y/X}$ unidades más arriba y otra paralela $2\hat{\sigma}_{Y/X}$ unidades más abajo, entonces, aproximadamente el 95% de los puntos del diagrama de dispersión caerán entre los valores $\pm 2\hat{\sigma}_{Y/X}$. Es decir, aproximadamente el 95% de los residuos son menores de $\pm 2\hat{\sigma}_{Y/X}$

iii) Si se traza una línea paralela a la línea de regresión $3\hat{\sigma}_{Y/X}$ unidades más arriba y otra paralela $3\hat{\sigma}_{Y/X}$ unidades más abajo, entonces, aproximadamente el 100% de los puntos del diagrama de dispersión caerán entre los valores $\pm 3\hat{\sigma}_{Y/X}$. Es decir, aproximadamente el 100% de los residuos son menores de $\pm 3\hat{\sigma}_{Y/X}$

EJEMPLO 13.2

Continuando con el ejemplo 13.1,

a) Calcule el error estándar de estimación.

b) Determine los residuales muestrales. ¿Qué porcentaje de estos residuales son menores que $\pm\hat{\sigma}_{Y/X}$?. Ilustre en una gráfica.

SOLUCION.

a) En el ejemplo 13.1 se han obtenido las siguientes estadísticas:

$$n = 10\,,\quad a = 6.38\,,\quad b = 0.6991\,,$$

$$\sum_{i=1}^{10} y_i = 384,\quad \sum_{i=1}^{10} x_i y_i = 19550\,,\quad \sum_{i=1}^{10} y_i^2 = 16{,}168\,,$$

Entonces, la varianza estimada o varianza de la regresión muestral es:

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-2} = \frac{16{,}168 - 6.38 \times 384 - 0.6991 \times 19550}{8}$$

$$s^2 = \frac{50.21}{8} = 6.276$$

El **error estándar de estimación** es: $\hat{\sigma}_{Y/X}$ o $s = \sqrt{6.276} = 2.51$

b) La tabla que sigue muestra los valores observados, los valores predecidos, y los residuales

Residuales

Observada X	Observada Y	Estimada $\hat{Y}$	Residual $Y - \hat{Y}$
28.0000	25.0000	25.95598	− 0.95598
25.0000	20.0000	23.85868	− 3.85868
35.0000	32.0000	30.84970	1.15030
40.0000	37.0000	34.34521	2.65479
45.0000	40.0000	37.84072	2.15928
50.0000	40.0000	41.33623	− 1.33623
50.0000	45.0000	41.33623	3.66377
35.0000	30.0000	30.84970	− 0.84970
70.0000	55.0000	55.31827	− 0.31827
80.0000	60.0000	62.30930	− 2.30930

Como se observa el 70% de los residuales de la muestra son menores que el error de estimación 2.51.

La gráfica 131.3 muestra la ubicación de los residuales con referencia a ±2.51

Distribución muestral de la estadística *b*

Se verifica (ver por ejemplo, referencia 1 página 334) que la estadística b se puede escribir como una combinación lineal de las variables aleatorias independientes Y_i . Esto es,

$$b = \sum_{i=1}^{n} c_i Y_i, \quad \text{donde, } c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Si se supone que cada variable Y_i tiene distribución normal, entonces, la estadística b tiene también distribución normal.

También, se verifica que **la media** de la estadística b, denotada por: μ_b o por $E(b)$, es igual a β, esto es:

$$E(b) = \beta$$

(la estadística b es una estimación insesgada del parámetro β).

Así mismo, se verifica que la **varianza** de b denotada por σ_b^2 o por $V(b)$ está dada por la expresión:

$$\sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

La desviación estándar o **error estándar** de b es el valor: $\sigma_b = \sqrt{\sigma_b^2}$

La variable aleatoria b tiene pues, distribución normal $N(\beta, \sigma_b^2)$ y la variable:

$$Z = \frac{b - \beta}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}}$$

se distribuye como normal $N(0,1)$.

Por otro lado, la variable aleatoria $(n-2)s^2/\sigma^2$ tiene distribución chi-cuadrado con $n-2$ grados de libertad. Además, b y s^2 son independientes.

Por consiguiente, la variable:

$$T = \frac{Z}{\sqrt{\dfrac{V}{n-2}}} = \frac{b - \beta}{\sqrt{\dfrac{s^2}{\sum(x_i - \bar{x}^2)}}} = \frac{b - \beta}{\hat{\sigma}_b}$$

tiene distribución t-student con $n-2$ grados de libertad.

La varianza de b es: $\hat{\sigma}_b^2 = \dfrac{s^2}{\sum x^2 - n(\bar{x}^2)} = \dfrac{6.276}{2807.6} = 0.002235$

El error estándar de b es : $\hat{\sigma}_b = \sqrt{\hat{\sigma}_b^2} = \sqrt{0.002235} = 0.0473$

En la tabla t-student con 8 grados de libertad y al nivel de confianza 0.95, se obtiene: $t_0 = t_{0.975,8} = 2.306$.

El intervalo de confianza del 95% para β se obtiene de los siguientes limites de confianza:

$$b \mp t_0\hat{\sigma}_b, \qquad 0.6991 \mp 2.306 \times 0.0473, \qquad 0.6991 \mp 0.109,$$

Luego, $\beta \in [0.5901, 0.8081]$, con confianza del 95%.

Como β=0 no pertenece al intervalo de confianza, debemos concluir que β≠0. Además, el valor positivo del limite de confianza inferior indica que a un nivel de confianza del 95% la pendiente: $\beta > 0$.

Prueba de hipótesis para β

Para probar la significación de la pendiente de la regresión muestral, las hipótesis nula y alternativa son respectivamente:

$$H_0: \beta = 0 \text{ y } H_1: \beta \neq 0$$

Si se supone verdadera la hipótesis nula, entonces, la estadística de la prueba es:

$$T = \frac{b}{\hat{\sigma}_b}$$

que tiene distribución t-student con $n-2$ grados de libertad.

Si $\beta \neq 0$, se puede realizar la prueba de la hipótesis nula: $H_0: \beta = \beta_0$ contra la hipótesis alternativa H_1 unilateral: $\beta < \beta_0$, o $\beta > \beta_0$ o bilateral: $\beta \neq \beta_0$

En este último caso, si se supone verdadera la hipótesis nula, entonces, la estadística $T = \dfrac{b - \beta_0}{\hat{\sigma}_b}$ tiene distribución t-student con $n-2$ grados de libertad.

EJEMPLO 13.4

Continuando con ejemplo 13.1, pruebe la significación de la pendiente de regresión muestral al nivel de significación $\alpha = 0.05$.

SOLUCION.

1. *Hipótesis*: $H_0: \beta = 0$ contra $H_1: \beta \neq 0$

2. *Nivel de significación*: $\alpha = 0.05$

3. *Estadística*: Se supone distribución normal. Si $H_0: \beta = 0$, es verdadera, la estadística de la prueba: $T = \frac{b}{\hat{\sigma}_b}$, tiene distribución es *t*-student con $n-2 = 10-2 = 8$ grados de libertad.

4. *Región crítica*: Con una prueba bilateral, con 8 grados de libertad y al nivel de significación del 0.05, en la tabla *t*-student se halla el valor crítico $t_{0.975,8} = 2.306$. Se rechazará H_0 si el valor calculado $t_k = b/\hat{\sigma}_b$ de la muestra es menor que – 2.306 o mayor que 0.2.306. Se aceptará H_0 en caso contrario.

5. *Cálculos*: $t_k = \frac{b}{\hat{\sigma}_b} = \frac{0.6991}{0.0473} = 14.78$

6. *Decisión*: Dado que 14.78 > 2.306, se rechaza H_0 lo que indica que existe una regresión significativa entre gastos e ingresos

NOTA. La significación bilateral (a dos colas) que resulta para esta prueba es la probabilidad: $P = 2P[t > 14.78]$=0.000 (paquete *MCEST*)

NOTA. El lector debería, como ejercicio, probar la hipótesis

$$H_0: \beta \leq 0.5 \text{ contra } H_1: \beta > 0.5.$$

Análisis de varianza para β

El análisis de varianza es un método que utiliza la estadística *F* para probar la significación de la ecuación de regresión muestral o la existencia de regresión en la población.

Es una prueba *F* de alternativa bilateral.

Las hipótesis nula y alternativa para esta prueba son respectivamente:

$$H_0: \beta = 0 \text{ contra } H_1: \beta \neq 0.$$

La estadística *F* de la prueba se obtiene de la partición de la varianza de *Y*: $S_Y^2 = \Sigma(y_i - \bar{y})^2/(n-1)$, en dos varianzas, la varianza no explicada y la varianza explicada por la regresión.

La partición, es la siguiente identidad de sumas de cuadrados:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

La primera suma: $\sum(y_i - \bar{y})^2$ se denomina *suma de cuadrados total* (*SCT*), refleja la variabilidad de los valores de Y con respecto a la media: $\bar{y}$.

La segunda suma: $\sum(y_i - \hat{y}_i)^2$ es la *suma de cuadrados de los errores* o *no explicada* (*SCE*).

La tercera suma se denomina *suma de cuadrados explicado por la regresión* (*SCR*), refleja la cantidad de variabilidad de los valores de Y explicada por la recta de regresión.

Las sumas de cuadrados: *SCT*, *SCE* y *SCR* tienen los grados de libertad respectivos: $n-1$, $n-2$, y 1.

Por otro lado, la distribución de SCR/σ^2 es $\chi^2(1)$ y la de SCE/σ^2 es $\chi^2(n-2)$, entonces, la variable aleatoria F definida por:

$$F = \frac{SCR/1}{SCE/(n-2)} = \frac{CMR}{CME} \text{ tiene distribución } F(1, n-2)$$

Dado el nivel de significación α, y los grados de libertad 1 y $n-2$, en la tabla de probabilidades F, se encuentra el valor crítico $c = F_{1-\alpha,1,n-2}$.

La regla de decisión consiste en rechazar la hipótesis nula: $H_0: \beta = 0$, si el valor calculado de F a partir de los datos de la muestra, es mayor que el valor crítico c. No se rechazará H_0 en caso contrario.

La prueba de la hipótesis nula $H_0: \beta = 0$ se resume en la siguiente tabla de análisis de varianza:

ANOVA para $\beta = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F calculada
Regresión	*SCR*	1	*CMR* = SCR/1	$F = \frac{CMR}{CME}$
Error	*SCE*	$n-2$	*CME* = SCE/($n-2$)	
Total	*SCT*	$n-1$		

Las sumas de cuadrados están dadas por:

$$SCT = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

$$SCR = b\left(\sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y})\right)$$

$$SCE = SCT - SCR$$

NOTA. Observe que la estimación de la varianza σ^2, se obtiene de la tabla ANOVA, pues, $s^2 = CME$.

EJEMPLO 13.5

Continuando con el ejemplo 13.1, utilice el método de análisis de varianza para probar la significación de la ecuación de regresión muestral, al nivel de significación 0.05.

SOLUCION.

1. *Hipótesis*: $H_0 : \beta = 0$ contra $H_1 : \beta \neq 0$
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística*: $F = \dfrac{CMR}{CME} \sim F(1, n-2)$, donde $n = 10$.
4. *Región crítica*: Para el nivel de significación $\alpha = 0.05$ y los grados de libertad 1, 8, en la tabla F se halla el valor crítico $F_{0.95,1,8} = 5.32$. Se rechazará H_0 si el valor calculado de F es mayor que 5.32. Se aceptará H_0 en caso contrario.
5. *Cálculos*: De los datos se obtiene:

$$SCT = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 16{,}168 - 10(38.4)^2 = 1{,}422.4$$

$$SCR = b\left(\sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y})\right) = 0.6991[19{,}550 - 10(45.8)(38.4)] = 1{,}372.19$$

$$SCE = SCT - SCR = 1{,}422.4 - 1{,}372.19 = 50.21.$$

La tabla de análisis de varianza es:

ANOVA para $\beta = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F calculada
Regresión	1372.19	1	$CMR = 1372.19$	$F = 218.5$
Error	50.21	8	$CME = 6.28$	
Total	1422.40	9		

6. *Decisión*: Dado que $F = \dfrac{1,372.198}{6.2875} = 218.5 > 5.32$, se debería rechazar H_0.

Estos resultados reflejan la validez del modelo de regresión poblacional entre ahorros e ingresos.

NOTA. La significación unilateral de esta prueba es, $P = P[F > 218.5] = 0.000$ (paquete *MCEST*)

NOTA. La estimación de la varianza σ^2 es $s^2 = CME = 6.28$.

13.1.5 Intervalos de estimación

Después de haber determinado que existe regresión lineal poblacional, entonces, se puede utilizar la ecuación lineal de regresión muestral, $\hat{Y} = a + bX$, para realizar predicciones o pronósticos válidos.

Las estimaciones están sujetas a una serie de variaciones que se pueden explicar mediante intervalos de estimación si el tamaño de la muestra es grande y la dispersión alrededor de la línea de regresión es aproximadamente normal

Hay dos tipos de intervalos de estimación

i) **Intervalo de confianza para la media de *Y* dado un valor de *X*,** es decir, estimar el valor μ_{Y/X_0} mediante un intervalo cuando $X = x_0$.

ii) **Intervalo de predicción para *Y* dado un valor de *X*,** es decir de Y_0, cuando $X = x_0$.

Intervalo de confianza para la media de Y: μ_{Y/X_0}

Sea μ_{Y/x_0} el valor de la media $\mu_{Y/X}$ cuando $X = x_0$.

Sea $\hat{y}_0$ el valor de $\hat{Y} = a + bX$, cuando, $X = x_0$. Esto es, $\hat{y}_0$ es un valor de la variable: $\hat{Y}_0 = a + b(x_0)$.

Para determinar el intervalo de confianza de μ_{Y/x_0} se utiliza la distribución muestral de la estadística: $\hat{Y}_0 = a + b(x_0)$.

La distribución muestral de la estadística $\hat{Y}_0$ es normal (ver por ejemplo, referencia 1 página 346):

Con media, $\mu_{\hat{Y}_0}$ o $E(\hat{Y}_0) = E(a + bx_0) = \alpha + \beta x_0 = \mu_{Y/x_0}$

y con varianza:

$$\sigma^2_{\hat{Y}_0} = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right], \text{ donde } S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x^2 - n(\bar{x})^2$$

La estimación de la varianza, $\sigma^2_{\hat{Y}_0}$, es la varianza:

$$\hat{\sigma}^2_{\hat{Y}_0} = s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]$$

La estadística:
$$T = \frac{\hat{Y} - \mu_{Y/x_0}}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}}$$

tiene distribución t-Student con $n - 2$ grados de libertad.

Observe $E(\hat{Y}_0) = \mu_{Y/X}$, es decir, que la estadística $\hat{Y}_0$ es un estimador puntual insesgado de μ_{Y/x_0}.

Luego, el intervalo de confianza del $(1 - \alpha)100\%$ para la media: μ_{Y/x_0} se obtiene, de los siguientes límites de confianza:

$$\hat{y}_0 \mp t_0\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

donde $t_0 = t_{n-2,1-\alpha/2}$ es un valor de t-Student con $n - 2$ grados de libertad

EJEMPLO 13.6

Continuando con el ejemplo 13.1 determine el intervalo de confianza del 95% para la media μ_{Y/x_0} cuando $x_0 = 90$

SOLUCION.

De la ecuación de regresión muestral cuando, $x_0 = 90$ se obtiene,

$$\hat{y}_0 = a + b(x_0) = 6.381 + 0.6991(90) = 69.3$$

Además, $n = 10$, $\bar{x} = 45.8$, $s = \sqrt{6.28} = 2.5$,

$S_{XX} = 2807.6$, $t_{0.975,8} = 2.306$

Luego, de los limites de confianza:

$$69.3 \mp 2.306(2.5)\sqrt{\frac{1}{10} + \frac{(90-45.8)^2}{2807.6}}\,, \quad 69.3 \mp 5.14\,.$$

se obtiene el intervalo de confianza, $64.16 \leq \mu_{Y/x_0} \leq 74.44$..

Esto es, si un empleado de la empresa tiene un salario semanal de $90, se estima que el gasto medio, se encuentre entre 64.16 y 74.44 dólares, con una confianza del 95%.

Intervalo de predicción para y_0

Para obtener un intervalo de predicción para un solo valor y_0 de la variable Y_0 $Y_0 = \alpha + \beta(x_0)$, se considera a la diferencia $\hat{y}_0 - y_0$ como un valor de la variable aleatoria $\hat{Y}_0 - Y_0$ cuya distribución muestral puede demostrase que es normal con **media cero** y desviación estándar estimada

$$\hat{\sigma}_{\hat{Y}_0 - Y_0} = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}}$$

El intervalo de predicción del $(1 - \alpha)100\%$ para el valor y_0, se determina utilizando la estadística:

$$T = \frac{\hat{Y} - Y_0}{s\sqrt{+1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}}$$

que tiene distribución es *t*-Student con $n-2$ grados de libertad.

El intervalo de predicción del $(1-\alpha)100\%$ para el sólo y_0, se obtiene de los límites de predicción:

$$\hat{y}_0 \mp t_0\, s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}}$$

en donde $t_0 = t_{n-2,1-\alpha/2}$, $\hat{y}_0$ es el valor de $\hat{Y}$ cuando $X = x_0$.

EJEMPLO 13.7

Continuando el ejemplo 13.1 determine el intervalo de predicción del 95% para y_0, cuando $x_0 = 90$

SOLUCION.

Cuando $x_0 = 90$, se obtiene

$$\hat{y}_0 = a + b(x_0) = 6.381 + 0.6991(90) = 69.3$$

Además, $n = 10$, $\bar{x} = 45.8$, $s = \sqrt{6.28} = 2.5$,

$$S_{XX} = 2807.6\ ,\ t_{0.975,8} = 2.306$$

El intervalo de predicción del 95% para y_0 se obtiene de los siguientes límites de predicción:

$$69.3 \mp 2.306(2.5)\sqrt{1+\frac{1}{10}+\frac{(90-45.8)^2}{2807.6}}\ ,$$

$$69.3 \mp 7.73$$

Luego, $61.57 \leq y_0 \leq 77.03.$

Esto es, si un empleado de la empresa tiene un salario semanal de \$90, se estima que el gasto semanal, se encuentre entre 61.57 y 77.03 dólares, con una confianza del 95%

Observe que este intervalo de predicción es más amplio que el se construyó para la media de *Y*.

13.1.6 Correlación

La **correlación lineal simple** es una técnica estadística que consiste en medir la relación lineal entre dos variables X e Y. La medida de la magnitud de la relación entre las dos variables se denomina *coeficiente o índice de correlación simple*.

Estas variables deben tener escala al menos de intervalo.

Los supuestos son los siguientes:

1. Tanto X como Y son variables aleatorias. No tienen que ser designadas como dependientes o independientes.
2. La poblacional bivariante es normal. Una población normal bivariante es aquella en la que X e Y están distribuidas normalmente.

   Las medias y varianzas respectivas son: μ_X , μ_Y y σ_X^2 , σ_Y^2 .
3. La relación entre X e Y es lineal.

El coeficiente de correlación lineal lo estudiaremos basándonos en el coeficiente de determinación.

Coeficiente de determinación

El coeficiente de determinación poblacional se denota por ρ^2 y se define en las dos formas siguientes:

$$\rho^2 = 1 - \frac{\sigma_{Y/X}^2}{\sigma_Y^2} = \beta^2 \frac{\sigma_X^2}{\sigma_Y^2}$$

en donde, β es la pendiente de la relación lineal entre X e Y,

$\sigma_{Y/X}^2$ denotada también por σ^2 es la varianza de la regresión lineal y.

σ_Y^2 es la varianza de Y.

Dado que $\sigma_{Y/X}^2 \leq \sigma_Y^2$, se tiene que: $0 \leq \rho^2 \leq 1$.

Además, $\rho^2 = 0$, cuando, $\beta = 0$ (cuando la recta de regresión es horizontal y por tanto no se puede predecir Y a partir de X).

Finalmente, $\rho^2 = 1$, cuando, $\sigma_{Y/X}^2 = 0$. (cuando hay relación lineal perfecta entre X e Y).

Coeficiente de determinación muestral.

Para n datos muestrales $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, puede obtenerse una estimación del coeficiente de determinación poblacional. Esta estimación o **coeficiente de determinación muestral** que se denota por r^2 o por R^2 se define por:

$$R^2 = b^2 \frac{s_X^2}{s_Y^2}$$

Esta última expresión es equivalente a: $R^2 = \dfrac{SCR}{SCT} = 1 - \dfrac{SCE}{SCT}$

Donde, las sumas de cuadrados: $SCT = SCE + SCR$ son respectivamente

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

El coeficiente de determinación R^2 se puede **interpretar** en tres formas:

1) *Como medida del mejoramiento en términos de reducción del error total.* Cuando $SCE = 0$, entonces $R^2 = 1$, y cuando $SCR = 0$, entonces, $R^2 = 0$. Luego, R^2 representa la reducción relativa en la suma de SCT cuando la recta de regresión se ha ajustado. Así,

   $R^2 = 1$ significa que hubo una reducción del 100% en la SCT o que $\hat{y}_i = y_i$ para todo i, y que todos los puntos están en la recta de regresión.

   $R^2 = 0$ indica que la reducción de SCT fue del 0% o que $\hat{y}_i = \overline{y}$ para cada i y que la línea de regresión es paralela al eje X.

   Observe que el 100% de la variación total es igual al $R^2 \times 100\%$ de la variación explicada por la recta de regresión más el porcentaje restante no explicado.

2) *Como la medida de la bondad del ajuste de la línea de regresión a los puntos.* Cuándo el ajuste es perfecto $R^2 = 1$. Cuando la línea es horizontal, $R^2 = 0$, es decir; no hay regresión lineal. Por lo tanto, *cuanto mayor sea el valor de R^2 mejor será el ajuste y más útil la recta de regresión como instrumento de predicción.*

3) Como medida de la linealidad de los puntos. Cuando R^2 se acerca a 1 la gráfica de los datos se acerca mucho a una línea recta. Si no es lineal la relación entre las dos variables, $R^2 = 0$.

NOTA. Cuando n es pequeño, el coeficiente de determinación R^2 es sesgado positivamente. Para corregir este sesgo, se calcula el **coeficiente de determinación ajustado** $\dot{R}^2$ o $\bar{R}^2$ que se define por:

$$\dot{R}^2 = 1 - \frac{CME}{CMT}$$

Donde, $CME = SCE/(n-2)$ y $CMT = SCT/(n-1)$.

Las sumas de cuadrados son las mismas del análisis de varianza.

Cuando se halla la ecuación de regresión, se aconseja calcular ambos coeficientes de determinación: R^2 y $\dot{R}^2$.

EJEMPLO 13.8.

Continuando con ejemplo 13.1 calcule el coeficiente de determinación R^2, y el coeficiente de determinación ajustado $\dot{R}^2$. Interprete los resultados.

SOLUCION.

Del ejemplo 13.1 resultan

$SCT = 1{,}422.4$, grado de libertad: $n-1 = 10-1 = 9$

$SCR = 1{,}372.19$, grado de libertad = 1

$SCE = SCT - SCR = 1{,}422.4 - 1{,}372.19 = 50.21$ **, grado de libertad = 8**

Entonces, $$R^2 \text{ o } r^2 = \frac{SCR}{SCT} = \frac{1{,}372.19}{1{,}422.4} = 0.965$$

El R^2 ajustado es, $$\dot{R}^2 = 1 - \frac{CME}{CMT} = 1 - \frac{50.21/8}{1{,}422.4/9} = 0.96$$

El valor de R^2 y el de $\dot{R}^2$ se interpretan en la misma forma. Por ejemplo, $\dot{R}^2 = 0.96$, significa que el 96% de la variación de la variable dependiente es explicada por la regresión lineal.

Coeficiente de correlación

El coeficiente de correlación lineal de Población se denota por ρ y se define como la raíz cuadrada del coeficiente de determinación poblacional ρ^2, esto es:

$$\rho = \beta \frac{\sigma_X}{\sigma_Y}$$

El coeficiente de correlación poblacional, ρ tiene el mismo signo aritmético que β la pendiente de la línea de regresión poblacional.

Dado que $0 \le \rho^2 \le 1$, se tiene, $-1 \le \rho \le 1$.

El valor de $\rho = 0$, cuando $\beta = 0$, es decir, no hay relación lineal cuando la recta de regresión es horizontal.

El valor de $\rho = \pm 1$, si, la varianza de la regresión poblacional σ^2 o $\sigma^2_{Y/X} = 0$, es decir, cuando hay relación lineal perfecta entre *X* e *Y*.

El valor de $\rho = +1$, significa que hay una relación lineal perfecta entre *X*, *Y*, con una pendiente positiva.

El valor de $\rho = -1$, significa que hay una relación lineal perfecta entre *X*, *Y*, con una pendiente negativa.

Coeficiente de correlación muestral

Cuando se extrae una muestra aleatoria de *n* pares de observaciones $(x_i, y_i), i = 1,2,...,n$ de la población (*X*, *Y*), se puede estimar el coeficiente de correlación ρ por

$$r = b\frac{S_X}{S_Y}$$

Esta estimación se denomina **coeficiente de correlación muestral** *r* **de Pearson. Se denota también por *R*.**

Se verifican las siguientes equivalencias:

$$r = \frac{Sxy}{S_X S_Y} = \frac{\sum xy - n(\bar{x})(\bar{y})}{\sqrt{[\sum x^2 - n(\bar{x})^2][\sum y^2 - n(\bar{y})^2]}}$$

donde: S_{XY} es la covarianza de *X* e *Y*,

S_X : es la desviación estándar de X,

S_Y : es la desviación estándar de Y

El coeficiente de correlación muestral r se interpreta igual que ρ.

NOTA. $r = 0$, implica una falta de linealidad y no de asociación entre X e Y, puede indicar por ejemplo, una relación cuadrática entre X e Y.

13.1.8. Inferencias acerca de ρ

El valor del coeficiente de correlación muestral, r está sujeto a variaciones muestrales.

El valor positivo o negativo de r, no implica necesariamente que el correspondiente valor de ρ sea positivo o negativo. Aun más, el valor de $r \neq 0$ no implica que el valor de $\rho \neq 0$.

Si el coeficiente de correlación de la muestra diferente de cero, ¿es el coeficiente de correlación en la población igual a cero?

Para resolver este problema debemos realizar una prueba de la significación del coeficiente de correlación muestral.

Las hipótesis nula y alternativa para esta prueba son respectivamente:

$$H_0 : \rho = 0 \text{ contra } H_1 : \rho \neq 0$$

Probar que $\rho = 0$ es equivalente a probar que $\beta = 0$ en la ecuación de regresión poblacional, ya que existe la relación, $\rho = \beta \dfrac{\sigma_X}{\sigma_Y}$.

Por lo tanto, si se rechaza (o no se rechaza) que: $\beta = 0$, entonces se rechaza (o no se rechaza) que: $\rho = 0$.

Sin embargo, cuando, se supone que $H_0 : \rho = 0$ es verdadera, la variable aleatoria,

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

tiene distribución t-student con $n - 2$ grados de libertad.

En consecuencia, se pude utilizar la estadística t-Student para probar:

$$H_0 : \rho = 0 \text{ contra } H_1 : \rho \neq 0$$

Se deja como ejercicio al lector verificar que:

$$t = \frac{b}{s/\sqrt{S_{XX}}} \text{ es equivalente a } t = r\sqrt{\frac{n-2}{1-r^2}}$$

EJEMPLO 13.9

Continuando con ejemplo 13.1, calcular el coeficiente de correlación r de Pearson y pruebe las hipótesis

$$H_0 : \rho = 0 \text{ contra } H_1 : \rho \neq 0$$

al nivel de significación $\alpha = 0.05$.

SOLUCION.

$$r = \frac{19550 - 10(4.58)(3.84)}{\sqrt{[23{,}784 - 10(4.58)^2][16{,}168 - 10(3.84)^2]}} = 0.98,$$

También, r se puede determinar utilizando la expresión $r = b\dfrac{S_X}{S_Y}$.

En la tabla t-student con 8 grados de libertad se halla el valor:

$$t_{n-2,1-\alpha/2} = t_{8,\,0.975} = 2.306,$$

La región crítica de la prueba es $RC = \{T < -2.306 \text{ o } T > 2.306\}$.

Además, $t = 0.98\sqrt{\dfrac{8}{1-(0.98)^2}} = 13.93$

Por lo tanto se rechaza la hipótesis de no asociación lineal.

NOTA. Cuando, $\rho \neq 0$,es decir, para hacer una prueba más general de $H_0 : \rho = \rho_0$ (supuesta verdadera) contra cualquier alternativa unilateral o bilateral, se utiliza la estadística :

$$Z = \frac{\sqrt{n-3}}{2}\ln\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$$

cuyo valor se compara con los [puntos críticos de la distribución normal estándar.

Para utilizar esta estadística se supone que *X* e *Y* siguen la distribución normal bivariada. En este caso, la cantidad:

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

es el valor de una variable aleatoria que se distribuye aproximadamente como una normal con media $(1/2)\ln[(1+\rho)/(1-\rho)]$ y varianza $1/(n-3)$, ver referencia 16, página 413.

EJEMPLO 13.10

Continuando con ejemplo 13.1, pruebe las hipótesis

$$H_0 : \rho = 0.9 \text{ contra } \quad H_1 : \rho > 0.9$$

al nivel de significación $\alpha = 0.05$.

SOLUCION.

Se supone que *X* e *Y* siguen la distribución normal bivariada.

Con α=0.05, la región crítica de la prueba es: RC={Z>1.645}.

El valor de *Z* obtenida de la muestra es:

$$Z = \frac{\sqrt{n-3}}{2}\ln\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right] = \frac{\sqrt{10-3}}{2}\ln\left[\frac{(1+0.98)(1-0.9)}{(1-0.98)(1+0.9)}\right] = 2.1836$$

Por lo que se concluye que es verdad que hay evidencia de que el coeficiente de correlación poblacional excede al valor 0.9.

Por otro lado, en la tabla normal estándar se obtiene la probabilidad

$$P = P[Z > 2.18] = 0.0146$$

Por lo que se concluye que la significación unilateral de la prueba es 0.0146.

EJERCICIOS

1. a) Ajuste a los puntos $(x_1, y_1),...,(x_n, y_n)$ una recta de la forma: $\hat{Y} = BX$, usando el "método de mínimos cuadrados"

   b) Determine la ecuación de la curva dada en a) si los puntos son:

   (2, 5), (4, 13), (5, 14), (6, 20), (7, 19), (10, 32)

   Rp. a) $B=\sum XY/\sum X^2$, b) $Y=3.065X$

2. Ajuste una recta de la forma $\hat{Y} = a + bX$ al conjunto de datos $(x_1, y_1),...,(x_n, y_n)$. Utilizando el método de mínimos cuadrados.

   Rp. $b=(n\sum xy - \sum x \sum y) / (n\sum x^2 - (\sum x)^2)$, $a = \bar{y} - b\bar{x}$

3. Verifique que el coeficiente de determinación r^2 es igual a $b^2 S_X^2 / S_Y^2$ donde b es la pendiente de la regresión, S_X^2 y S_Y^2 son las varianzas de X e Y respectivamente.

4. Si $(x_1, y_1),...,(x_n, y_n)$ son n pares de datos observados que se encuentran en la recta L: $\hat{Y} = a + bX$, ¿qué porcentaje de la varianza total de los y_i es explicado por L?.

   Rp. 100%, por que $\sum(y - \hat{y}_i)^2=0$.

5. Dada la recta de regresión de mínimos cuadrados $\hat{Y} = a + bX$, si se produce un incremento igual a c en uno de los valores de X, ¿cuánto es el incremento respectivo que se produce en $\hat{Y}$?.

   Rp. bc.

6. Al realizar la regresión de Y en X basado en una muestra aleatoria de 10 pares de datos (x_i, y_i), se tiene que la varianza de los y_i es igual a 16.5 y que la suma de cuadrados debido a la regresión es 155. ¿Qué porcentaje de la varianza de los y_i es explicada por la regresión?.

   Rp. R^2= SCR/SCT=155/165=0.9394.

7. Se tiene la siguiente información: $(x_1, y_1),...,(x_n, y_n)$ relativa a X: ingresos, e Y: egresos.

Para predecir el egreso, un investigador usa la media $\overline{y}$ para cada x_i y mide el error total de predicción con $\sum_{i=1}^{n}(y_i-\overline{y})^2$, resultando 60. En cambio, otro investigador predice el egreso usando la recta de mínimos cuadrados y mide el error total de predicción con $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ resultando 3.

Halle el coeficiente de correlación entre ingresos y egresos sabiendo que cuando aumenta el ingreso aumenta el egreso.

Rp. SCT=SCE+SCR, 60=3+SCR, SCR=57, r^2=SCR/SCT=57/60=0.95, r=0.9746.

8. Si n pares $(x_1,y_1),\ldots,(x_n,y_n)$ tiene índice de correlación r comprobar que la recta de regresión para los puntos $(x_1^*,y_1^*),\ldots,(x_n^*,y_n^*)$, en donde $x_i^*=\dfrac{x_i-\overline{x}}{s_X}$, $y_i^*=\dfrac{y_i-\overline{y}}{s_Y}$ para $i=1,2,\ldots,n$, es $y^*=rx^*$.

Rp. $\overline{x^*}=0$, $\overline{y^*}=0$, $r^*=r$, entonces, $y^*=rx^*$

9. El gerente de personal de la empresa P&C quiere estudiar la relación entre el ausentismo y la edad de sus trabajadores. Tomó una muestra aleatoria de 10 trabajadores de la empresa y encontró los siguientes datos.

Edad (años)	25	46	58	37	55	32	41	50	23	60
Ausentismo (días por año)	18	12	8	15	10	13	7	9	16	6

a) Use el método de mínimos cuadrados para hallar la ecuación muestral que relaciona las dos variables.

b) Calcule el coeficiente de determinación. De su comentario sobre el ajuste de la línea de regresión a los datos de la muestra.

c) Calcule el error estándar de la estimación $\hat{\sigma}_{Y/X}$ y los residuales. ¿Qué porcentaje de residuales de la muestra son menores de $\pm\ \hat{\sigma}_{Y/X}$?. ¿Qué opina del ajuste por este método?

d) Determine si el coeficiente de regresión en la población es diferente de cero si se sabe que el error estándar del coeficiente de regresión muestral es 0.056. Use el nivel de significación 0.01

Rp. a) $\hat{Y}=22.946-0.26X$, b) 0.728 mal ajuste, c) 2.246, según *MCEST*, 8 de los 10 residuales en la muestra están dentro de un error estándar de la línea de regresión. Un buen ajuste para una muestra pequeña, d) $t_{calc}=-4.628$, signifc=0.002

10. El Banco "PRESTAMO" estudia la relación entre las variables, ingresos (X) y ahorros (Y) mensuales de sus clientes. Una muestra aleatoria de sus clientes reveló los siguientes datos en dólares:

X	350	400	450	500	950	850	700	900	600
Y	100	110	130	160	350	350	250	320	130

a) ¿Cuáles son los supuestos del modelo de regresión?
b) Dibuje el diagrama de dispersión y describa la tendencia trazando una línea a través de los puntos
c) Determine la ecuación de regresión muestral. Interprete esta ecuación
d) Calcule el error estándar de estimación. ¿Entre qué dos valores estarán aproximadamente 95% de las predicciones? (Suponga muestra grande)
e) Analice que tan bien se ajustan los puntos del diagrama de dispersión a la línea de regresión utilizando el coeficiente de determinación.

Rp. c) $\hat{Y} = -74.92 + 0.452X$, d) σ=30.85, $\hat{Y} \pm 2\times\sigma$, e) $r^2 = 0.93$ el 93% de la varianza de Y es explicada por la regresión.

11. Continuando con el ejercicio 10,
a) Calcule la desviación estándar $\hat{\sigma}_b$ de la pendiente b (error estándar de b)
b) Halle un intervalo de confianza de 0.95 para β. ¿Se puede afirmar que β=0?.
c) Utilice la prueba t bilateral para probar la hipótesis nula H_0: $\beta = 0$ al nivel de significación del 5%. Calcule la probabilidad P.

Rp. a) $\hat{\sigma}_b = 0.0482$, b) [0.34, 0.56], no, c) $t = 9.37$, gl =7, P=0.000, se rechaza H_0: $\beta = 0$, la ecuación de regresión lineal muestral es significativa.

12. Continuando con el ejercicio 10, la pendiente de la línea de regresión muestal resultó, $b = 0.452$, se quiere determinar si esta pendiente es significativa en la población utilizando el método de análisis de varianza.
a) Plantee las hipótesis nula y alternativa
b) Determine la región de rechazo al nivel de significación 0.05 y describa la regla de decisión
c) Describa la tabla ANOVA y tome la decisión.
d) Halle la probabilidad P de la prueba

Rp. a) H_0: $\beta = 0$ y H_1:$\beta \neq 0$, b) $F(1,7)$, RC=]5.59, +∞[, c) SCR = 83626.05, SCE = 6662.82, SCT = 90288.87, F=87.86, se rechaza H_0 existe regresión lineal poblacional, d) 1.1328E−05.

14. Continuando con el ejercicio 10 determine el intervalo de confianza del 95% para;
a) La cantidad de ahorro promedio μ_{Y/x_0}, si el ingreso es x_0 = \$1200.

b) La cantidad de ahorro y_0, cuando el ingreso es $x_0 = \$1200$.

Rp. a) [398, 536], b) [366.6, 567.5]

15. Continuando con el ejercicio 10
    a) Calcule el coeficiente de correlación. Interprete la tendencia.
    b) ¿Por qué son iguales los signos de b y r ?.
    c) Utilizando la significación al 5% del coeficiente regresión muestral, ¿podemos concluir que hay relación positiva entre ahorros e ingresos?
    d) Realice la prueba bilateral de la hipótesis nula $H_0: \rho = 0$ al nivel de significación 0.05.

Rp. a) $r = 0.96$ tendencia positiva, b) por que $r=b(S_X/S_Y)$ en la población $\rho=\beta(\sigma_X/\sigma_Y)$, c) se rechaza H_0: $\beta=0$, entonces, se rechaza $\rho=0$. Si

16. Un comerciante mayorista encargó un estudio para determinar la relación entre los gastos de publicidad semanal por radio y las ventas de sus productos. En el estudio se obtuvieron los siguientes resultados

Semana	2	3	4	5	6	7	8	9	10	11
Gastos de publicidad ($)	30	20	40	30	50	70	60	80	70	80
Ventas ($)	300	250	400	–	550	750	630	930	700	840

En la quinta semana por diversos motivos no se pudo hacer el estudio.

a) Determine la ecuación de regresión de ventas sobre gastos de publicidad.
b) Interprete la pendiente de regresión.
c) ¿En cuánto estimaría las ventas de la quinta semana?. ¿Cuánto es el error o residual en esa semana?

Rp. a) $\hat{Y} = -4.07 + 10.77X$, b) si los gastos de publicidad aumentan \$1, las ventas aumentan en promedio \$10.77, c) $\hat{y} = \$319.03$, no se puede determinar.

17. Continuando con el ejercicio 16,
    a) Haga el análisis de la validez de la regresión poblacional de ventas sobre gastos de publicidad al nivel del 5% por medio de un ANOVA.
    b) Halle el coeficiente de determinación y el coeficiente de determinación ajustado. De su comentario sobre estos valores.
    c) A partir del análisis realizado en a) ¿qué puede usted concluir acerca de la correlación entre ventas y gastos de publicidad?.

Rp. a) $SCR = 443618.7$, $SCE = 11003.5$, $SCT = 454622.2$, $F = 282.2$, gl = 1, 7, signific=0.000, se rechaza H_0: $\beta = 0$, b) $R^2 = 0.976$, R^2ajust =0.972 existe un buen ajuste, c) $r = 0.988$ es significativo.

18. Se obtuvieron los siguientes datos para determinar la relación entre cantidad de fertilizante y producción de papa por hectárea.

Sacos de Fertilizante por hectárea	3	4	5	6	7	8	9	10	11	12
Rendimiento en Quintales	45	48	52	55	60	65	68	70	74	76

a) Encuentre la ecuación de regresión de la cosecha sobre el fertilizante, por el método de mínimos cuadrados.

b) Estime la cosecha si se aplican 12 sacos de fertilizantes. ¿Cuánto es el error o residual?

c) Determine el coeficiente de determinación. De su comentario sobre este valor.

d) Determine si el coeficiente de regresión muestral es significativo utilizando el método de intervalo de confianza para β al nivel de confianza 0.95. Es posible concluir que

e) Desarrolle un intervalo del 95% para el rendimiento promedio de papa que se obtendría si se utiliza 12 kilogramos de fertilizante.

f) Desarrolle un intervalo de predicción del 95% para el rendimiento de papa que se obtendría si se utiliza 12 kilogramos de fertilizante.

Rp. a) $\hat{Y} = 34.255 + 3.606X$, b) $\hat{y} = 77.58$, 1.58, c) $R^2 = 0.991$, d) ES(b)=0.119, IC=[3.332, 3.88], $\beta \neq 0$, e) Int media, IC=[76.07, 78.99], f) Int predicc=[74.64, 80.41].

19. El número de horas de estudio invertidas y las calificaciones finales en un curso de Matemáticas de una muestra 10 alumnos ha dado los siguientes resultados:

Alumno	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
Horas de estudio	14	16	22	20	18	16	18	22	10	08
Calificación	12	13	15	15	17	11	14	16	08	05

a) Determine la recta de regresión de la calificación sobre el número de horas de estudio invertidos. Interprete la ecuación de regresión

b) Use el método de la prueba t para probar la hipótesis nula H_0: $\beta = 0$ contra una alternativa bilateral. Utilice el valor P en la conclusión.

c) Calcule el grado de asociación entre calificación y horas de estudio, ¿es significativo al nivel 1%?.

d) Halle el error estándar de estimación. $\hat{\sigma}_{Y/X}$. ¿Qué porcentaje de los residuales de la muestra son menores que $\pm\ \hat{\sigma}_{Y/X}$? . Realice una ilustración gráfica

Rp. a) $\hat{Y} = 0.565 + 0.734X$, b) ES(b)=0.111, GL=8, t=6.613, Prob P=0.000, se rechaza H_0: β=0 c) R=0.919,se rechaza H_0:ρ =0 al 1%, d) s=1.56, Residuos: 1.16, 0.69, −1.71, −0.24, 3.23, −1.31, 0.23, −0.71, 0.097, −1.44, 80% un porcentaje alto.

20. Sobre la base de una muestra de tamaño 28 se encontró que la ecuación de regresión muestral de gastos mensuales (Y) sobre tamaño de la familia (X) es:

$$\hat{Y} = 3.975 + 0.563X$$

Además la covarianza de Y con X es igual a 32, y la desviación estándar de Y es igual a 5,

a) Determine el coeficiente de correlación y analizar la bondad del ajuste de la línea de regresión con el coeficiente de determinación.
b) Desarrolle la tabla ANOVA para determinar si el coeficiente de regresión poblacional es distinto de cero al nivel de significación del 5%.
c) ¿Es la correlación muestral significativa al 1%?.

Rp. a) S_X=7.539, r=bS_X/S_Y=0.84889, R^2=0.72, b) SCT=675, SCR=486, SCE=189 F=66.859, gl=1,26, signific.=0.000, se rechaza H_0: β=0, c) Se rechaza ρ = 0. a cualquier nivel.

21. Una muestra de 60 de las 350 agencias de ventas de automóviles de una importadora registradas en un mes con X (autos vendidos por agencia), Y(ventas en miles de dólares) ha dado los siguientes resultados:

$$\bar{x}=10,\quad \bar{y}=20,\quad \sum x^2=7{,}000,\quad \sum y^2=42{,}000,\quad \sum xy=8{,}000$$

a) Determine la ecuación de regresión: $\hat{Y}=a+bX$.
b) Calcule el coeficiente de determinación. ¿Qué porcentaje de la variación total es explicada por la regresión?.
c) Pruebe la hipótesis nula H_0: β = 0 contra una alternativa bilateral al nivel de significación 0.05. Enuncie las suposiciones necesarias.
d) Pruebe la hipótesis nula H_0: ρ = 0 contra una alternativa bilateral

Rp. a) $\hat{Y}$ = 60 − 4X, b) r = −0.94, r^2 = 0.8836, el 88.36%. de la varianza de Y es explicada por la regresión., c) SCR =15904.8, SCE =20.95.2, SCT =18,000 F = 440.2, gl = 1,58, se rechaza H_0: β = 0, d) en consecuencia se rechaza ρ = 0.

22. Los contadores con frecuencia estiman los gastos generales basándose en el nivel de producción. En la tabla que sigue se da la información recabada sobre gastos generales y las unidades producidas en 10 plantas y se desea estimar una ecuación de regresión para estimar gastos generales futuros:

Gastos generales ($)	300	1000	1100	1200	600	800	900	500	400	200
Unidades producidas	15	45	55	75	30	40	45	20	18	10

a) Determine la ecuación de regresión y haga un análisis de los coeficientes de regresión
b) ¿Proporcionan los datos suficiente evidencia para indicar que las unidades producidas aportan información para predecir los gastos generales?.
c) Realice un análisis de la bondad del ajuste de la ecuación de regresión lineal.
d) ¿Qué puede usted concluir acerca de la correlación poblacional entre gastos generales y unidades producidas?.

Rp. a) $\hat{Y}$ = 116.652 + 16.525X, b) SCR =1034492.25, SCE =65507.75, SCT = 1100000, F=126.335, gl=1,8, $P[F>126.335]$=0.000, se re rechaza H_0: β =0, c) R^2=0.94, el ajuste es bueno, d) Se rechaza que ρ = 0, existe correlación

23. Las calificaciones de un grupo de estudiantes en el examen parcial (X) y en el examen final (Y), fueron las siguientes:

X	Y	X	Y	X	Y	X	Y
12	15	18	20	15	17	13	14
08	10	12	14	12	15	10	13
10	12	10	12	11	12	12	15
13	14	12	10	12	13	13	14
09	12	14	16	11	12	12	13
14	15	09	11	10	13	16	18
11	16	12	13	14	12	15	17

a) Determine la ecuación de regresión lineal de Y en X
b) Pruebe la significación de la pendiente poblacional en $\alpha = 5\%$.
c) Calcule el error estándar de la pendiente b y halle el intervalo de confianza del 95% para β. ¿Se puede aceptar que $\beta=0.9$?
d) Halle R, R^2, R^2 ajustado. Interprete los resultados
e) ¿Es significativo el coeficiente de correlación muestral?

Rp. a) $\hat{Y} = 3.469 + 0.856X$, b) SCR =100.583, SCE =50.845, SCT =151.429 F =51.434, GL=1,26, signific=0.000 se rechaza H_0: $\beta=0$, c) $ES(b)$=0.119, IC=[0.61, 1.101], si, d) R=0.815, R^2=0.664, R^2ajust=0.651., e) Si

24. Con los siguientes datos muestrales:

Coeficiente de inteligencia: IQ	135	115	95	100	110	120	125	130	140
Notas de un examen	16	13	12	12	14	14	15	15	18

a) Halle la ecuación de regresión muestral.
b) Interprete la pendiente de regresión parcial.
c) Utilizando t-Student pruebe la hipótesis $\beta = 0$, contra la hipótesis $\beta > 0$ al nivel de significación $\alpha=0.05$. ¿Se puede aceptar que $\beta=1$?
d) Determine el grado de asociación entre las dos variables.
e) Utilizando t-Student pruebe la hipótesis $\rho = 0$ contra la hipótesis $\rho > 0$ al nivel de significación $\alpha=0.05$.

Rp. a) $\hat{Y} = 0.276 + 0.118\,X$, c) $ES(b)$=0.016, GL=7, t=7.171, signific.=0.000, d) r = 0.938, e) t=6.2, rechaza H_0: $\rho = 0$.

25. En un estudio para determinar la relación entre edad (X) y presión sanguínea (Y) una muestra aleatoria de 9 mujeres ha dado los siguientes resultados:

X	54	40	70	35	62	45	55	50	38
Y	148	123	155	115	150	126	152	144	114

a) Encuentre la ecuación de regresión de Y en X y estime la presión sanguínea para una mujer de 75 años.

b) Utilizando t-Student pruebe la hipótesis $\beta = 0.9$, contra la hipótesis $\beta > 0.9$ al nivel de significación α=0.05

c) Pruebe la hipótesis nula H_0: $\rho = 0.9$ contra H_1: $\rho > 0.9$

Rp. a) $\hat{Y} = 69.96 + 1.33X$, b) $ES(b)$=0.201, t=(1.33−0.9)/ES=2.139, gl =7, RC={T>1.86}, se rechaza H_0: β=0.9, hay evidencia que β es mayor que 0.9, c) r=0.929, z=0.33, se acepta ρ=0.9 .

26. Se estudia la relación entre la variable: X = Nivel socioeconómico con sus categorías: Bajo=0, Medio=1, Alto=2 y la variable: Y = Gastos en educación quincenales (en dólares). Una muestra aleatoria de 27 familias ha dado los siguientes datos:

X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
0	64	1	80	2	95	0	70
1	69	2	89	0	77	1	65
2	94	0	84	1	56	2	90
0	55	1	82	2	80	0	64
1	60	2	99	0	50	1	67
2	92	0	73	1	50	2	80
0	70	1	76	2	89		

a) Determine la ecuación de regresión de gastos sobre ingresos.

b) ¿Proporcionan los datos suficiente evidencia para indicar que el nivel socioeconómico aporta información para predecir los gastos en educación?.

c) Realice un análisis de la bondad del ajuste de la ecuación de regresión lineal utilizando el coeficiente de determinación.

d) A partir del análisis realizado en b), ¿qué puede usted concluir acerca de la correlación entre nivel socioeconómico y gastos en educación?.

Rp. a) $\hat{Y} = 63.648 + 11.167X$, b) SCR=2244.5, SCE=2939.57, SCT=5184.07, F=19.089, gl =1, 25, significac=0.000, se rechaza H_0: $\beta = 0$, c) R^2=0.433, el ajuste no es bueno, d) Se rechaza H_0: $\rho = 0$..

27. Las cantidades de un compuesto químico (Y) que se disuelven en 100 gramos de agua a diferentes temperaturas (X) se registraron en la tabla que sigue:

X (°C)	Y gramos				
0	10	8	10	9	11
15	15	12	14	16	18
30	27	23	25	24	26
45	33	30	32	35	34
60	46	40	43	42	45
75	50	52	53	54	55

a) Encuentre la ecuación de regresión de Y en X

b) Estime la varianza de la regresión poblacional.

c) Determine el coeficiente de regresión estandarizado beta

d) Calcule el error estándar de la pendiente *b*. Además desarrolle un intervalo de confianza del 95% para β. ¿Se puede aceptar que β=0.6?.

e) Determine un intervalo de confianza del 95% para la cantidad promedio de producto químico que se disolverá en 100 gramos de agua a 50° C.

f) Determine un intervalo de predicción del 95% para la cantidad de producto químico que se disolverá en 100 gramos de agua a 50° C.

Rp. a) $\hat{Y} = 7.705 + 0.587X$, b) 4.723, c) 0.99, d) $ES(b)$=0.15, IC de β: [0.556, 0.619], si e) ES=0.44, IC=[36.17, 37.98], f) ES=2.218, IC= [32.53, 41.62].

28. El gerente de ventas de una cadena de tiendas obtuvo información (ver tabla que sigue) de los pedidos por internet y del número de ventas realizadas por esa modalidad. Como parte de su presentación en la próxima reunión de vendedores al gerente le gustaría dar información específica sobre la relación entre el número de pedidos y el número de ventas realizadas.

Tienda	1	2	3	4	5	6	7	8	9	10
Número de pedidos	50	56	60	68	65	50	79	35	42	15
Número de ventas	45	55	50	65	60	40	75	30	38	12

a) Use el método de mínimos cuadrados para expresar a relación entre estas dos variables.

b) Haga un análisis de los coeficientes de regresión

c) ¿Proporcionan los datos suficiente evidencia para indicar que las unidades producidas aportan información para predecir los gastos generales?.

d) Realice un análisis de la bondad del ajuste de la ecuación de regresión lineal.

e) ¿Qué puede usted concluir acerca de la correlación poblacional entre gastos generales y unidades producidas?.

Rp. a) $\hat{Y} = 116.652 + 16.525X$, b) SCR =1034492.25, SCE =65507.75, SCT = 1100000, F=126.335, gl =1,8, $P[F>126.335]$=0.000, se re rechaza H_0: β = 0, c) r^2 = 0.94, el ajuste es bueno, d) Se rechaza H_0: β = 0, en consecuencia se rechaza que ρ = 0, existe correlación poblacional.

28. Con los datos de la hoja de cálculo del *estudio socioeconómico de los estudiantes universitarios de Lima* (ver apéndice) realice una análisis de regresión simple eligiendo la variable X_8 como variable dependiente, y la variable X_6 como variable independiente.

13.2 Regresión lineal múltiple.

13.2.1 Modelo de regresión lineal múltiple

El análisis de regresión múltiple es una técnica estadística que consiste en la extensión del análisis de regresión simple a aplicaciones que implican dos o más variables independientes o **predictoras:** $X_1, X_2, \ldots, X_k$ $(k \geq 2)$ para pronosticar el valor de la variable dependiente *Y*.

La variable dependiente *Y* es de escala de intervalo o de razón

Es una técnica muy útil empleada en diversas disciplinas, como por ejemplo, en economía y finanzas.

Con la aplicación de paquetes de computo se hace posible la solución de problemas en las que intervienen un gran número de variables.

En el modelo de la regresión lineal múltiple para *k* variables independientes $X_1, X_2, \ldots, X_k$ la media de *Y* se expresa por la ecuación :

$$\mu_{Y/X_1,X_2,\ldots,X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

en donde: $\beta_0, \beta_1, \ldots, \beta_k$ son los **coeficientes de regresión poblacional**.

La estimación de la ecuación de regresión poblacional es la ecuación de regresión lineal múltiple muestral cuya expresión es:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

en donde: $b_0, b_1, \ldots, b_k$ son los **coeficientes de regresión muestral**.

Cada coeficiente de regresión poblacional β_i se estima mediante el respectivo coeficiente de regresión muestral b_i, utilizando el método de mínimos cuadrados.

Para $k = 2$ la gráfica de la ecuación de regresión es un *plano* que intercepta a *Y* en β_0. Los demás coeficientes de regresión parcial β_i son las pendientes de la línea de regresión de *Y* con la variable X_i mientras las otras variables independientes se mantienen constantes.

Estas pendientes indican *el cambio promedio de Y correspondiente a un incremento unitario en X_i cuando las demás X permanecen constantes.*

Para *k* variables independientes ($k > 2$), la gráfica de la ecuación de regresión poblacional es un *hiperplano* en el espacio de $k + 1$ dimensiones.

Para visualizar la relación entre la variable dependiente y cada una de las variables independientes se pude utilizar diagramas de dispersión.

Los **supuestos en el análisis de regresión múltiple** son los mismos de la regresión lineal simple. En particular para hacer inferencias acerca de los parámetros β_i se debe suponer que la variable dependiente Y es normal con media μ_{Y/X_i} y varianza σ^2

13.2.2. Determinación de la ecuación de regresión muestral

Los coeficientes de regresión muestral $b_0, b_1, \ldots, b_k$ se calculan a partir de los datos de una muestra aleatoria. Los datos de la muestra aleatoria de tamaño n se pueden recopilar en la forma:

$$(x_{1i}.x_{2i}, \ldots, x_{ki}, y_i)\ , \quad i = 1, 2, \ldots, n \ \text{ y } \ n > k$$

en donde y_i es la respuesta observada (valor de la variable dependiente Y) para los valores $x_{1i}.x_{2i}, \ldots, x_{ki}$ de las k variables independientes respectivas $X_1, X_2, \ldots, X_k$.

Para cada $i = 1, 2, \ldots, n$ los datos de la muestra satisfacen la ecuación de regresión poblacional:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

donde, ε_i, el término *error*, es una variable aleatoria que se supone tiene media 0 y varianza σ^2.

Para hacer inferencias acerca de los parámetros β_i se debe suponer que la variable ε_i es normal con media μ_{Y/X_i} y varianza σ^2.

Los datos de la muestra satisfacen también, la ecuación de regresión muestral:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

donde $e_i = y_i - \hat{y}_i$ es el término *residual*.

Los coeficientes de regresión muestral $b_0, b_1, b_2, \ldots, b_k$ se calculan por **el método de mínimos cuadrados.**

Este método consiste en determinar los coeficientes de manera que hagan mínima la suma de los cuadrados de los residuales (*SCE*) expresada por:

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2$$

Derivando *SCE* cada vez con respecto a $b_0, b_1, b_2, \ldots, b_k$ e igualando a cero, se obtienen las $k+1$ **ecuaciones normales** que siguen:

$$\begin{array}{l}
nb_0 + b_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_k \sum x_k = \sum y \\
b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \cdots + b_k \sum x_1 x_k = \sum x_1 y \\
b_0 \sum x_2 + b_1 \sum x_2 x_1 + b_2 \sum x_2^2 + \cdots + b_k \sum x_2 x_k = \sum x_2 y \\
\vdots \\
b_0 \sum x_k + b_1 \sum x_k x_1 + b_2 \sum x_k x_2 + \cdots + b_k \sum x_k^2 = \sum x_k y
\end{array}$$

donde, $\sum x_j = \sum_{i=1}^{n} x_{ji}$, para $j = 1, 2, \cdots, k$.

Estas ecuaciones pueden ser resueltas para $b_0, b_1, b_2, \ldots, b_k$ por cualquier método apropiado para resolver sistemas de ecuaciones lineales.

Sin embargo, existen muchos paquetes de computo disponibles para llevar a cabo estos cálculos, entre ellos, el paquete de computo estadístico didáctico *MCEST* elaborado por el autor de este texto.

NOTA. En el modelo de regresión lineal múltiple, las variables independientes pueden incluir variables cualitativas (o categóricas), como por ejemplo, sexo. En este caso, los valores de la variable llamada también **variable indicadora,** deben ser codificadas.

La codificación de la variable sexo por ejemplo, puede ser 0=mujer, 1=hombre (o viceversa).

EJEMPLO 13.10.

Se realiza un estudio de asociación entre las siguientes variables:

Y : Gastos mensuales expresados en cientos de dólares
X_1 : Ingreso mensual familiar en miles de dólares
X_2 : Tamaño de la familia.

En una muestra de 10 familias escogidas al azar se han encontrado los datos que se presentan en la tabla que sigue:

Y	X_1	X_2
45	10	9
40	9	8
38	8	6
35	7	6
32	7	5
30	6	4
28	6	3
27	4	2
25	3	2
22	2	1

a) Determine la ecuación de regresión muestral de los gastos mensuales con respecto a las dos variables: Ingreso mensual y número de hijos.

b) Estime el gasto mensual para una familia de 8 hijos y cuyo ingreso mensual es $7,000.

SOLUCION.

a) La ecuación de regresión múltiple muestral a determinar es:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 .$$

De los datos de la muestra, resultan:

$$n = 10, \ \sum x_1 = 62, \ \sum x_2 = 46, \ \sum x_1^2 = 444, \ \sum x_1 x_2 = 345$$

$$\sum x_2^2 = 276, \quad \sum y = 322, \ \sum x_1 y = 2158, \ \sum x_2 y = 1653$$

El sistema de ecuaciones normales de mínimos cuadrados es:

$$10b_0 + 62b_1 + 46b_2 = 322$$
$$62b_0 + 444b_1 + 345b_2 = 2158$$
$$46b_0 + 345b_1 + 276b_2 = 1653$$

Las soluciones únicas de este sistema son:

$$b_0 = 18.947, \quad b_1 = 0.509, \quad b_2 = 2.195$$

Por lo tanto, la ecuación de regresión múltiple muestral es:

$$\hat{Y} = 18.947 + 0.509X_1 + 2.195X_2$$

b) Si una familia tiene 8 hijos y si su ingreso mensual es \$7,000, esto es, si $x_1 = 7$, $x_2 = 8$, entonces, sus gasto mensual estimado es:

$$\hat{y} = 18.947 + 0.509 \times 7 + 2.195 \times 8 = 40.07 \text{ o } \$4{,}007$$

NOTA. Un **método práctico** para convertir el sistema de 3 ecuaciones a un sistema de dos ecuaciones, consiste en escribir:

$$b_1 S_{X_1X_1} + b_2 S_{X_1X_2} = S_{X_1Y}$$

$$b_1 S_{X_1X_2} + b_2 S_{X_2X_2} = S_{X_2Y}$$

en donde: $S_{X_1X_1} = \sum x_1^2 - n(\bar{x}_1)^2$, $S_{X_2X_2} = \sum x_2^2 - n(\bar{x}_2)^2$,

$S_{X_1X_2} = \sum x_1x_2 - n(\bar{x}_1)(\bar{x}_2)$, $S_{X_1Y} = \sum x_1 y - n(\bar{x}_1)(\bar{y})$

$S_{X_2Y} = \sum x_2 y - n(\bar{x}_2)(\bar{y})$

Del sistema reducido se obtienen b_1 y b_2, y luego se obtiene,

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

EJEMPLO 13.11.

Continuando con el ejemplo 13.10, obtenga los coeficientes de regresión muestral utilizando el método del sistema reducido

SOLUCION.

Con los datos del ejemplo 3.10 se obtiene

$$S_{X_1X_1} = \sum x_1^2 - n(\bar{x}_1)^2 = 444 - 10 \times (6.2)^2 = 59.6$$

$$S_{X_2X_2} = \sum x_2^2 - n(\bar{x}_2)^2 = 276 - 10 \times (4.6)^2 = 64.4$$

$$S_{X_1X_2} = \sum x_1x_2 - n(\bar{x}_1)(\bar{x}_2) = 345 - 10 \times 6.2 \times 4.6 = 59.8$$

$$S_{X_1Y} = \sum x_1 y - n(\bar{x}_1)(\bar{y}) = 2158 - 10 \times 6.2 \times 32.2 = 161.6$$

$$S_{X_2Y} = \sum x_2 y - n(\bar{x}_2)(\bar{y}) = 1653 - 10 \times 4.6 \times 32.2 = 171.8$$

Entonces,

$$b_1 59.6 + b_2 59.8 = 161.6$$

$$b_1 59.8 + b_2 64.4 = 171.8$$

de donde resultan: $b_1 = 0.509$, $b_2 = 2.195$, y

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = 32.2 - 0.509 \times 6.2 - 2.195 \times 4.6 = 0.1894$$

Interpretación de la ecuación de regresión múltiple

Cuando se tiene una ecuación de regresión múltiple donde las unidades de medición son las mismas tanto para la variable dependiente *Y*, como para las variables independientes, $X_1, X_2, \ldots, X_k$, los coeficiente de regresión parcial se comparan directamente en la siguiente forma:

La cantidad b_0 es la *ordenada en el origen*. Para $k = 2$ por ejemplo, es el intercepto del plano de regresión con el eje *Y* cuando $X_1 = 0$ y $X_2 = 0$.

Los demás coeficientes de regresión parcial b_i *indican el cambio promedio de Y correspondiente a un incremento unitario en* X_i *cuando las demás X permanecen constantes.*

Por **ejemplo**, en la ecuación de regresión múltiple del ejemplo 13.10:

$$\hat{Y} = 18.947 + 0.509X_1 + 2.195X_2$$

La cantidad $b_0 = 18.947$ es la *ordenada en el origen.*

El valor $b_1 = 0.509$ indica que si hay un aumento unitario en X_1 es decir un aumento de \$1,000 en los ingresos mensuales de las familia, los gastos mensuales (*Y*) aumentan en promedio 0.509 en cientos de dólares, esto es, 0.509×100 = \$50.9, manteniendo constante X_2 (el número de hijos de la familia).

El valor $b_2 = 2.195$ indica que si hay un aumento de un miembro en la familia (X_2) entonces hay un aumento promedio de \$2.195 cientos de dólares, 2.195×100= \$219.5 en los gastos mensuales, permaneciendo constante X_1 (ingresos mensuales).

Esto se debe a que las unidades de b_1 y b_2 son respectivamente:

$$b_1 = \frac{\text{Gastos (en cientos de dólares)}}{\text{Ingresos (en miles de dólares)}}$$

$$b_2 = \frac{\text{Gastos (en cientos de dólares)}}{\text{Número de miembros de la familia}}$$

Coeficientes de regresión Beta

Cuando se tiene una ecuación de regresión muestral múltiple donde las unidades de medición no son las mismas para la variable dependiente Y, y para las variables independientes $X_1, X_2, \ldots, X_k$, los coeficiente de regresión parcial no se pueden comparar directamente. Para superar esta dificultad **se utilizan los coeficientes de regresión estandarizados beta.**

Las unidades de medición de las variables Y, $X_1, X_2, \ldots, X_k$ *se transforman a unidades de desviaciones estándares dividiendo cada variable por su desviación estándar.*

Por **ejemplo**, en la ecuación de regresión muestral:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2,$$

se tiene:
$$\frac{\hat{Y}}{s_Y} = \frac{b_0}{s_Y} + (b_1 \frac{s_{X_1}}{s_Y}) \frac{X_1}{s_{X_1}} + (b_2 \frac{s_{X_2}}{s_Y}) \frac{X_2}{s_{X_2}}$$

Los coeficientes definidos por:

$$\text{beta}_i = b_i \frac{s_{X_i}}{s_Y}$$

son denominados ***coeficientes beta*** o ***coeficientes de regresión parcial estándar***.

Los coeficientes beta se interpretan como sigue: "Si hay una variación de una desviación estándar en X_i habrá una variación de beta_i desviaciones estándar en la variable Y.

Continuando con el ejemplo 3.10, la variable X_2 tiene diferente unidad de medición de Y. Además, $s_{X_2} = 2.675$, $s_Y = 7.239$, resultando el coeficiente $\text{beta}_2 = 2.195 \times \frac{2.675}{7.239} = 0.811$, luego, cuando hay una variación de una desviación estándar en X_2 habrá una variación de 0.811 desviaciones estándares en Y. Esto es, si hay un aumento de 2.675 en el número de hijos en la familia, habrá un aumento en los gastos mensuales de $0.811 \times 7.239 = 5.871$ en cientos de dólares.

Análogamente el coeficiente: $\text{beta}_1 = 0.181$.

13.2.3. Pruebas de significación de los coeficientes de regresión.

Una vez determinada la ecuación de regresión muestral debemos determinar si los coeficientes de esa ecuación de regresión son significativos o no. Es decir, se debe determinar si los coeficientes de regresión calculados a partir de la muestra implican que los correspondientes coeficientes de regresión poblacional son o no son distintos de cero.

Si todos los coeficientes de regresión poblacional son iguales a cero no podremos predecir Y, es decir no habría regresión lineal. Si sólo uno de ellos es igual a cero, por ejemplo, $\beta_2 = 0$; podemos concluir que no hay regresión de Y en la variable X_2.

El análisis de la regresión debería comenzar con una **prueba global de significación** de los coeficientes de regresión muestral mediante un análisis de varianza ANOVA. Si se acepta que no todos los coeficientes de regresión poblacional son iguales a cero, entonces, se debe analizar la significación de los coeficientes de regresión muestral en forma individual.

Para la prueba de significación de los coeficientes de regresión se requiere suponer que la variable dependiente Y es normal con varianza σ^2.

El proceso es el siguiente:

1) Análisis de varianza. Prueba global de los coeficientes de regresión

El análisis de varianza se utiliza en este caso para determinar si existe o no regresión lineal en la población de la variable dependiente Y con **todas** las variables independientes *en conjunto* (**análisis global** de los coeficientes de regresión). Las hipótesis nula y alternativa de la prueba son respectivamente:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{al menos una de las } \beta_i \text{ es distinto de cero}$$

La estadística F de la prueba de la hipótesis nula H_0 contra H_1 se obtiene de la partición de varianza de Y en varianza residual (no explicada) y varianza explicada por la regresión lineal, cuyas sumas de cuadrados respectivos son:

$$\sum_{i=1}^{n}(y_i-\bar{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$SCT \quad = \quad SCE \quad + \quad SCR$$

Donde, *SCT* es la suma de cuadrados total, *SCE* es la suma de cuadrados de errores o **residuales** (varianza no explicada) y *SCR* es la suma de cuadrados de la regresión (varianza explicada por la regresión).

Los grados de libertad respectivos de las sumas de cuadrados son:

$$n-1=(n-k-1)+k$$

Por otro lado, la estadística SCR/σ^2 se distribuye según una chi-cuadrado con k grados de libertad, esto es:

$$SCR/\sigma^2 \sim \chi^2(k).$$

También, la estadística:

$$SCE/\sigma^2 \sim \chi^2(n-k-1),$$

Estas dos estadísticas son independientes. Por tanto, el cociente de las dos chi^2 divididos entre sus respectivos grados de libertad, tiene distribución F con grados de libertad k y $n-k-1$, esto es,

$$F=\frac{(SCR/\sigma^2)/k}{(SCE/\sigma^2)/(n-k+1)}=\frac{SCR/k}{SCE/(n-k+1)}=\frac{CMR}{CME}\sim F(k,n-k-1)$$

en donde;

$$CMR=SCR/k \quad \text{y} \quad CME=SCE/(n-k-1)$$

son los cuadrados medios de regresión y de error respectivamente.

Dado el nivel de significación α, para los grados de libertad k, y $n-k-1$, en la tabla F se encuentra el valor crítico $c=F_{1-\alpha,k,n-k-1}$.

La región crítica de la prueba es el intervalo: $RC=\{F>c\}$.

La regla de decisión es: Rechazar, $H_0: \beta_1=\beta_2=\cdots=\beta_k=0$, si el valor calculado de F es mayor que el valor crítico c. No rechazar H_0 en caso contrario.

La prueba de hipótesis de análisis global se resume en la siguiente tabla de análisis de varianza (ANOVA):

ANOVA para $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F calculada
Regresión	SCR	k	$CMR = SCR/k$	$F = \dfrac{CMR}{CME}$
Error o residual	SCE	$n-k-1$	$CME = SCE/(n-k-1)$	
Total	SCT	$n-1$		

Las sumas de cuadrados SCT, SCR y SCE se calculan utilizando las siguientes expresiones:

$$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\overline{y})^2$$

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = b_1 S_{X_1Y} + b_2 S_{X_2Y},$$

donde $S_{X_1Y} = \sum_{i=1}^{n} x_1 y - n(\overline{x}_1)(\overline{y})$, y $S_{X_2Y} = \sum_{i=1}^{n} x_2 y - n(\overline{x}_2)(\overline{y})$

$$SCE = SCT - SCR$$

EJEMPLO 13.11b.

Continuando con el ejemplo 13.10 y utilizando el nivel de significación 0.05, investigue si alguna de las variables independientes tiene un coeficiente de regresión significativo.

SOLUCION.

La hipótesis nula y alternativa de esta prueba ANOVA o prueba global son:

$$H_0 : \beta_1 = \beta_2 = 0$$

H_1 : al menos una de las β_i es distinto de cero

De los datos del ejemplo 13.10, resultan:

$$SCT = \sum_{i=1}^{n} y_i^2 - n(\overline{y})^2 = 10{,}840 - 10 \times (32.2)^2 = 471.6$$

$$S_{X_1Y} = \sum_{i=1}^{n} x_1 y - n(\overline{x}_1)(\overline{y}) = 2158 - 10 \times 6.2 \times 32.2 = 161.6$$

$$S_{X_2Y} = \sum_{i=1}^{n} x_2 y - n(\overline{x}_2)(\overline{y}) = 1{,}653 - 10 \times 4.6 \times 32.2 = 171.8$$

$$SCR = b_1 S_{X_1Y} + b_2 S_{X_2Y} = 0.509 \times 161.6 + 2.195 \times 171.8 = 459.365$$

$$SCE = SCT - SCR = 471.6 - 459.365 = 12.235.$$

Los grados de libertad de *SCR*, *SCE* y *SCT* son respectivamente, 2, 7 y 9.

$$CMR = \frac{459.365}{2} = 229.6825 \qquad CME = \frac{12.235}{7} = 1.74786,$$

$$F = \frac{CMR}{CME} = \frac{229.683}{1.748} = 131.409$$

Las sumas de cuadrados, los grados de libertad, los cuadrados medios y la estadística *F* se resumen en la siguiente tabla de análisis de varianza.

ANOVA para $H_0 : \beta_1 = \beta_2 = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*F* Calculada
Regresión	459.365	2	229.683	*F* =131.409
Residual	12.235	7	1.748	
Total	471.600	9		

Al nivel de significación 5%, y con los grados de libertad 2 y 7, en la tabla de probabilidades de la *F* se encuentra el valor crítico $F_{0.95,\,2,\,7} = 4.74$. Dado que el valor calculado $F = 131.409 > 4.74$, debemos rechazar $H_0 : \beta_1 = \beta_2 = 0$. En consecuencia, de esta **prueba global**, podemos inferir que por lo menos uno de los coeficientes de regresión poblacional es diferente de cero y que existe regresión global de *Y* con X_1 y X_2.

Observe que la probabilidad *P* es $P = P[F(2,7) > 131.409] = 0.000$

NOTA. Si se decide aceptar la hipótesis nula H_0, se concluye que no hay regresión de *Y* globalmente con X_1 y X_2, y el análisis termina. Pero si se rechaza H_0 y aceptamos H_1 (al menos una de las β_i es distinto de cero) se debe continuar con el análisis, determinando las variables independientes que influyen en la regresión.

2) Prueba individual de los coeficientes de regresión

Si existe regresión de la variable dependiente *Y* globalmente con **todas** las variables independientes *X* *en conjunto*. Es deseable determinar **qué variables contribuyen en forma significativa** al modelo de regresión múltiple. Si alguna

variable independiente X_i no contribuye en forma significativa al modelo, se la debe descartar y buscar la ecuación lineal adecuada.

Para seleccionar el modelo de regresión adecuado, se pueden utilizar los siguientes métodos:

i) **Intervalo de confianza** del $(1-\alpha)100\,\%$ para cada coeficiente de regresión β_i. Si el intervalo de confianza respectivo, contiene el valor cero, se infiere que $\beta_i = 0$ y en consecuencia la variable X_i no contribuye significativamente al modelo de regresión lineal múltiple y se la descarta..

Se encuentra luego la ecuación de regresión con las variables independientes no descartadas y se repite el análisis, hasta encontrar la ecuación de regresión lineal satisfactoria.

ii) **Prueba de hipótesis t-student bilateral** para docimar

$$H_0 : \beta_i = 0 \text{ contra } H_1 : \beta_i \neq 0$$

Si la decisión es aceptar $H_0 : \beta_i = 0$ a un nivel de significación dado, entonces, la variable X_i no contribuye significativamente al modelo de regresión múltiple y debe ser descartada.

Se encuentra luego la ecuación de regresión con las variables independientes no descartadas y se repite el análisis, hasta encontrar la ecuación de regresión lineal satisfactoria

Los métodos de intervalo de confianza y la prueba de hipótesis *t* se explican en la sección 13.2.7.

iii) **Análisis de regresión por pasos.** Puede ser

a) De **selección hacia adelante:** Consiste en añadir a la ecuación de regresión lineal simple una variable independiente en cada uno de los pasos o etapas hasta llegar a obtener la ecuación de regresión satisfactoria.

b) De **eliminación hacia atrás**: Se comienza con la ecuación de la regresión que incluya todas las variables independientes. Si el modelo global es significativo, se selecciona una variable candidata a ser eliminada. Si la variable seleccionada no es significativa, se la elimina del modelo. Luego, se ajusta un modelo de regresión utilizando las variables independientes restantes y se repite el proceso hasta hallar la ecuación de regresión adecuada. La estadística es la *F*.

EJEMPLO 13.11c. Continuando con el ejemplo 13.10 aplique el método de eliminación por pasos hacia atrás para investigar la significación de los coeficientes individuales de regresión.

SOLUCION

En el ejemplo 13.10, se concluye que la ecuación de regresión muestral: $\hat{Y} = 18.947 + 0.509X_1 + 2.195X_2$ es significativo globalmente.

El coeficiente de X_1 en regresión lineal simple de Y con sólo X_1, que denotaremos por $b_1^{'}$, está dado por:

$$b_1^{'} = \frac{\sum x_1 y - n(\bar{x}_1)(\bar{y})}{\sum x_1^2 - n(\bar{x}_1)^2} = \frac{2158 - 10 \times 6.2 \times 32.2}{444 - 10 \times (6.2)^2} = \frac{161.6}{59.6} = 2.711$$

Igualmente, el coeficiente $b_2^{'}$ de X_2 en regresión lineal simple de Y con sólo X_2, está dado por:

$$b_2^{'} = \frac{\sum x_2 y - n(\bar{x}_2)(\bar{y})}{\sum x_2^2 - n(\bar{x}_2)^2} = \frac{1653 - 10 \times 4.6 \times 32.2}{276 - 10 \times (4.6)^2} = \frac{171.8}{64.4} = 2.668$$

La **suma de cuadrados de la regresión (explicada) debido solo a** X_1, sin considerar X_2 denotada por $SCR(X_1)$ o $R(\beta_1)$ es la expresión:

$$SCR(X_1) = b_1^{'}(\sum x_1 y - n(\bar{x}_1)(\bar{y})) = 2.711 \times 161.6 = 438.098$$

De igual manera, **la suma de cuadrados de la regresión (explicada) debido sólo** a X_2 sin considerar X_1, denotada por $SCR(X_2)$ o $R(\beta_2)$ es:

$$SCR(X_2) = b_2^{'}(\sum x_2 y - n(\bar{x}_2)(\bar{y})) = 2.668 \times 171.8 = 458.362$$

La **suma de cuadrados de la regresión de X_2 ajustada a X_1** denotada por $R(\beta_2/\beta_1)$ es:

$$R(\beta_2/\beta_1) = SCR - SCR(X_1) = 459.365 - 438.098 = 21.267$$

La **suma de cuadrados de la regresión de X_1 ajustada a X_2** denotada por $R(\beta_1/\beta_2)$ es:

$$R(\beta_1/\beta_2) = SCR - SCR(X_2) = 459.365 - 458.362 = 1.003$$

Se selecciona como variable a ser eliminada a la variable que da el valor más pequeño de la suma de cuadrados ajustada a la otra variable. En este caso es la variable X_1., cuya suma de cuadrados ajustada a X_2 es:

$$R(\beta_1/\beta_2) = 1.003$$

Se prueba entonces, $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$ utilizando la estadística F:

$$F = \frac{R(\beta_1 / \beta_2)}{CME} \text{ que se distribuye según la } F(1, n-k-1)$$

Se rechazará H_0 al nivel de significación α, si la F calculada de los datos de la muestra es mayor al valor crítico $F_{1-\alpha,1,n-k-1}$.

Se eliminará la variable X_1 del modelo de regresión, si no se rechaza H_0.

En este ejemplo, el valor crítico de la prueba es $F_{0.95,1,7} = 5.59$. La estadística calculada

$$F_{cal} = \frac{R(\beta_1 / \beta_2)}{CME} = \frac{1.003}{1.748} = 0.5738$$

Por lo tanto no se rechaza H_0. Se elimina pues, la variable X_1 del modelo.

La ecuación de regresión muestral de Y con X_2, (eliminando X_1) es:

$$\hat{Y} = 19.929 + 2.668X_2$$

El valor crítico de la prueba de la contribución de X_2 es $F_{0.95,1,8} = 5.32$, dado que 275.908 > 5.32 se concluye que el modelo resultante es significativo.

El proceso se resume en la siguiente tabla

Tabla ANOVA de eliminación por pasos hacia atrás

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F Calculada	Significac
Regres (X_1, X_2) Error Total	459.356 12.235 471.600	2 7 9	229.683 1.748	F = 131.409	0.000
Regresión (X_2) Error Total	458.311 13.289 471.600	1 8 9	458.311 1.661	275.908	0.000

NOTA. Este procedimiento de selección del modelo de regresión resultante por el método de eliminación por pasos hacia atrás, se desarrolla para un número grande de variables independientes utilizando un paquete de computo estadístico como por ejemplo el SPSS.

NOTA. En cada paso el valor de *CME* (varianza s^2) utilizada en la prueba F es el cuadrado medio del error para el modelo de regresión en esa etapa.

13.2.4. Coeficiente de determinación múltiple

Una vez hallada la ecuación de regresión muestral debería interesarnos conocer la bondad de ajuste a los datos de la muestra..

Para $k = 2$ variables independientes el plano ajustado a los puntos de la muestra que sea horizontal y pase por la media $(\bar{y}, \bar{x}_1, \bar{x}_2)$ puede considerarse como un plano básico con respecto al cual se mide la mejora introducida por la regresión.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

De la partición de la suma total de cuadrados en suma de cuadrados no explicada y suma de cuadrados explicada por la regresión, se obtiene el coeficiente de determinación múltiple que se denota por R^2 o por $R^2_{Y,12...k}$ donde k es el número de variables independientes.

El **coeficiente de determinación múltiple** se define en forma similar al coeficiente de determinación simple, por ejemplo, para k=2 variables independientes, se define por:

$$R^2_{Y.12} = 1 - \frac{SCE}{SCT}$$

La suma total de cuadrados contiene las desviaciones de los puntos observados Y de un plano que se ajusta a estos puntos y que es horizontal pasando por el punto $(\bar{y}, \bar{x}_1, \bar{x}_2)$ y a partir del cual se mide el mejoramiento producido por la regresión.

Al igual que el coeficiente de determinación R^2, el coeficiente de determinación múltiple mide el porcentaje de la varianza de Y que queda explicada al conocer dos o más variables independientes. Cuanto mayor es el valor de $R^2_{Y.12}$ menor es la dispersión y mayor el ajuste del plano de regresión a los datos.

Por **ejemplo**, continuando con el ejemplo 13.10 el coeficiente de determinación múltiple es:

$$R^2_{Y.12} = 1 - \frac{SCE}{SCT} = 1 - \frac{12.235}{471.6} = 0.974$$

Esto indica que aproximadamente el 97% de la varianza de los gastos mensuales (Y) queda estadísticamente explicado por los ingresos familiares (X_1) y por el número de hijos de las familias (X_2).

Cuando el tamaño n de la muestra es pequeño, el índice de determinación múltiple $R^2_{Y.12}$ tiende a estar positivamente sesgada.

Para corregir este sesgo se utiliza el *coeficiente o índice de determinación múltiple corregido* (*o ajustado*) que se denota por $\dot{R}^2$ o por $\dot{R}^2_{Y,12}$ y se define por:

$$\dot{R}^2_{Y.12} = 1 - \frac{CME}{CMT}$$

Por **ejemplo**, aplicando a los datos del ejemplo 13.10 el coeficiente de determinación múltiple corregido es:

$$\dot{R}^2_{Y.12} = 1 - \frac{CME}{CMT} = 1 - \frac{1.748}{52.4} = 0.967$$

La raíz cuadrada positiva del coeficiente de determinación múltiple se denomina **coeficiente de correlación múltiple** que denotamos por R o $R_{Y,12}$.

Este número mide la relación entre las variables independientes consideradas como grupo y la variable dependiente Y.

La prueba de la significación del coeficiente correlación múltiple poblacional es la misma prueba F que se utiliza para determinar si existe regresión global de Y con respecto a todas las variables independientes en conjunto.

Es decir, si se acepta $H_0: \mu_1 = \mu_2 = ... = \mu_k = 0$, entonces, se concluye que no existe correlación múltiple de Y con todas las variables independientes $X_1, X_2, ..., X_k$.

Por **ejemplo**, continuando con el ejemplo 13.10, el coeficiente de correlación múltiple de Y con X_1 y X_2 es:

$$R = \sqrt{R^2_{Y,12}} = \sqrt{0.974} = 0.987$$

Dado que en el análisis de varianza se encuentra que es significativa la regresión global de Y con X_1 y X_2, se concluye que existe correlación lineal múltiple en la población.

13.2.5. Modelo de regresión lineal mediante matrices

Supóngase que se tiene k variables independientes $X_1, X_2, \ldots, X_k$ y una variable dependiente Y, y que además la muestra aleatoria de tamaño n: $(x_{1i}.x_{2i}, \ldots, x_{ki}, y_i)$, $i = 1, 2, \ldots, n$ y $n > k$, satisface la ecuación de regresión poblacional:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta \; x_{ki} + \varepsilon_i$$

Con la notación matricial estas n ecuaciones se pueden escribir como el modelo poblacional:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

donde,

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

El vector de parámetros $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$, se estima por el vector $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$

Las entradas del vector β son los coeficientes de regresión poblacional y las del vector **b** son los coeficientes de regresión muestral.

El vector **b** se encuentra por el método de mínimos cuadrados, para lo cual se minimiza la suma de cuadrados de errores:

$$SCE = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$$

El sistema de ecuaciones normales se obtiene de la ecuación de derivadas en vectores:

$$\frac{\partial}{\partial \mathbf{b}}(SCE) = \mathbf{0}$$

No se presentarán aquí los detalles del método de mínimos cuadrados para llegar a la ecuación normal en forma vectorial. En las aplicaciones nos interesa saber que el vector **b** es la solución de la ecuación en matrices:

$$(\mathbf{X'X})\mathbf{b} = \mathbf{X'Y}$$

donde,

$$\mathbf{X'X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}, \qquad \mathbf{X'}Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix},$$

Si la matriz **X'X** es no singular, se puede escribir la solución para los coeficientes de regresión como

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Las técnicas para invertir la matriz **X'X** de orden $k+1$ por $k+1$ se explican en la mayor parte de los libros textos sobre determinantes y matrices elementales. Hay muchos paquetes de computadora para problemas de regresión múltiple entre ellos el *MCEST*, paquete que no sólo proporcionan los coeficientes de regresión muestral sino que proporciona abundante información sobre inferencias relativas a la ecuación de regresión poblacional.

EJEMPLO 13.12

Se realizó un estudio de las relaciones entre las notas obtenidas en 4 cursos por 10 estudiantes de administración de empresas seleccionados aleatoriamente. Las 3 variables independientes X_1, X_2, X_3 y la variable dependiente Y son :

Y = notas en estadística aplicada
X_1 = nota en matemática básica
X_2 = nota en lógica
X_3 = nota en economía general

Los datos se dan en la tabla que sigue.

a) Determine las ecuaciones normales $(\mathbf{X'X})\mathbf{b} = \mathbf{X'Y}$

b) Determine el vector solución de las ecuaciones normales: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ y estime el modelo de regresión lineal múltiple.

Y	X_1	X_2	X_3
13	12	10	18
15	14	12	15
17	16	10	18
16	15	11	20
13	11	13	15
17	15	16	18
19	16	10	19
14	14	14	16
18	17	13	15
11	11	11	13

SOLUCION.

Utilizando el paquete estadístico didáctico *MCEST* se obtiene:

a) Las ecuaciones normales $(\mathbf{X'X})\mathbf{b} = \mathbf{X'Y}$, cuya expresión es:

$$\begin{bmatrix} 10 & 141 & 120 & 167 \\ 141 & 2029 & 1694 & 2375 \\ 120 & 1694 & 1476 & 1995 \\ 167 & 2375 & 1995 & 2833 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 153 \\ 2203 \\ 1838 \\ 2584 \end{bmatrix}$$

b) La solución de la ecuación normales, el vector: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$, donde,

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 13.809 & -0.120 & -0.430 & -0.411 \\ -0.120 & 0.033 & -0.006 & -0.016 \\ -0.430 & -0.006 & 0.030 & 0.009 \\ -0.411 & -0.016 & 0.009 & 0.032 \end{bmatrix}^{-1} \begin{bmatrix} 153 \\ 2203 \\ 1838 \\ 2584 \end{bmatrix} = \begin{bmatrix} -2.91069 \\ 1.01726 \\ 0.04826 \\ 0.19690 \end{bmatrix}$$

Entonces

$$b_0 = -2.91069, \quad b_1 = 1.01726, \quad b_2 = 0.04826, \quad b_3 = 0.19690,$$

La ecuación de regresión muestral es:

$$\hat{Y} = -2.91069 + 1.01726X_1 + 0.04826X_2 + 0.1969X_3.$$

13.2.6. Estimadores de mínimos cuadrados

Si se supone que los errores aleatorios $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ son independientes cada uno con media cero y varianza σ^2, entonces, se puede demostrar que:

1) $b_0, b_1, \cdots, b_k$ son, respectivamente, estimadores insesgados de los coeficientes de regresión $\beta_0, \beta_1, \cdots, \beta_k$, estro es:

$$E(b_i) = \beta_i \quad i = 0, 1\,2, \ldots, k\,.$$

2) Los elementos de la matriz $\sigma^2(\mathbf{X'X})^{-1}$ representan las varianzas de los coeficientes de regresión muestral, $b_0, b_1, \cdots, b_k$; sobre la diagonal principal, y las covarianzas fuera de la diagonal.
Por ejemplo, para $k = 2$ variables independientes, se tiene:

$$\sigma^2(\mathbf{X'X})^{-1} = \sigma^2 \begin{bmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{bmatrix}$$

de donde para $i = 0, 1, 2,$ se obtiene:

$$\text{Var}(b_i) \ \text{ o } \ \sigma^2_{b_i} = c_{ii}\sigma^2, \ \text{ y}$$

$$\text{Cov}(b_i, b_j) = c_{ij}\sigma^2 \ \text{ para } i \neq j$$

3) Una estimación insesgada de la varianza σ^2 es la varianza muestral que denotaremos por s^2 o por $\hat{\sigma}^2$ y que se define por:

$$\hat{\sigma}^2 = \frac{SCE}{n-k-1} = CME$$

donde, $SCE = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

El **error estándar de estimación múltiple** es $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

4) La estimación de la varianza $\text{Var}(b_i)$ o $\sigma^2_{b_i} = c_{ii}\sigma^2$ es entonces,

$$\hat{\sigma}^2_{b_i} = c_{ii}\hat{\sigma}^2\,.$$

El error estándar de cada b_i es $\hat{\sigma}_{b_i} = \sqrt{c_{ii}\hat{\sigma}^2}$.

5) Para probar hipótesis y determinar intervalos de confianza de β_i $(i = 1, 2, \cdots, k)$ se debe suponer además, que los residuos ε_i tienen distribución normal. En este caso, los coeficientes b_i tienen también distribución normal con media β_i y varianza $c_{ii}\sigma^2$.

EJEMPLO 13.13.

Continuando con el ejemplo 13.12

a) Calcule $\hat{\sigma}^2$ o s^2, la estimación insesgada de la varianza de la regresión múltiple poblacional σ^2.

b) ¿Cómo interpreta el error estándar de la estimación múltiple?

c) Halle la varianza del vector **b**, Var(**b**) y el error estándar de cada uno de los coeficientes de la regresión muestral

SOLUCION

a) $\hat{\sigma}^2 = \dfrac{SCE}{n-k-1} = CME = \dfrac{5.824}{10-3-1} = 0.971,$ donde $SCE = SCT - SCE$

$$SCT = \sum y^2 - n(\overline{x})^2 = 2399 - 10(15.3)^2 = 58.1 \text{ y } SCR = \sum_{i=1}^{3} b_i S_{X_i Y} = 52.276$$

b) El **error estándar de la estimación** múltiple es, $\hat{\sigma} = \sqrt{\sigma^2} = 0.9852$

Se mide en las mismas unidades de la variable dependiente. Si los errores están distribuidos normalmente y si la muestra es grande, cerca del 68% de los residuales son menores que ±0.9852, cerca del 95% de los residuales son menores que ±2×0.9852 y cerca del 100% de los residuales son menores que ±3×0.9852.

c) $$\hat{V}(\mathbf{b}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = 0.971 \begin{bmatrix} 13.809 & -0.120 & -0.430 & -0.411 \\ -0.120 & 0.033 & -0.006 & -0.016 \\ -0.430 & -0.006 & 0.030 & 0.009 \\ -0.411 & -0.016 & 0.009 & 0.032 \end{bmatrix}^{-1}$$

$$\hat{V}(\mathbf{b}) = \begin{bmatrix} 13.40466 & -0.11611 & -0.41693 & -0.39924 \\ -0.11611 & 0.03188 & -0.00573 & -0.01584 \\ -0.41693 & -0.00573 & 0.02944 & 0.00865 \\ -0.39924 & -0.01584 & 0.00865 & 0.03107 \end{bmatrix}$$

de donde resultan las varianzas de los coeficientes de regresión:

$$\sigma^2_{b_0} = 13.40466,\quad \sigma^2_{b_1} = 0.03188,\quad \sigma^2_{b_2} = 0.02944,\quad \sigma^2_{b_3} = 0.03107$$

y los respectivos **errores estándares**:

$$\sigma_{b_0} = 3.66124,\quad \sigma_{b_1} = 0.17854,\quad \sigma_{b_2} = 0.17159,\quad \sigma_{b_3} = 0.17627.$$

13.2.7 Intervalos de predicción

Después de haber decidido que existe regresión lineal múltiple poblacional o que son significativos los coeficientes de regresión muestral podemos utilizar el modelo para:

i) Predecir la media $\mu_{Y/X_1,X_2,\ldots,X_k}$, dado los valores: $x_{10}, x_{20}, \cdots, x_{k0}$. y determinar el intervalo de confianza de la **respuesta promedio,** o

ii) Predecir una nueva observación de Y dado los valores: $x_{10}, x_{20}, \cdots, x_{k0}$ de $X_1, X_2, \ldots, X_k$ y determinar el intervalo de predicción de la **respuesta observada.**

Como en el caso de la regresión lineal simple, se **hace la suposición adicional de que los errores (o residuales) son independientes y tienen distribución normal.**

Intervalo de confianza para $\mu_{Y/X_{10},X_{20},\ldots,X_{k0}}$

Sea $\mu_{Y/x_{10},x_{20},\ldots,x_{k0}}$ el valor de la media $\mu_{Y/X_1,X_2,\ldots,X_k}$, dado los valores: $x_{10}, x_{20}, \cdots, x_{k0}$.

Y sea $\hat{y}_0$ el valor de, $\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i$, en el valor $x_{10}, x_{20}, \cdots, x_{k0}$. Esto es:

$$\hat{y}_0 = b_0 + \sum_{i=1}^{k} b_i x_{i0}\ ).$$

Para determinar el intervalo de confianza de la media $\mu_{Y/x_{10},x_{20},\ldots,x_{k0}}$ se utiliza la distribución muestral de, $\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i$.

La distribución muestral de la estadística $\hat{Y}$ es normal (ver por ejemplo, referencia 1 página 346) con :

media: $\mu_{Y/x_{10},x_{20},\ldots,x_{k0}}$

y varianza: $\sigma_{\hat{Y}}^2 = \sigma^2 \mathbf{x}_0'(\mathbf{X'X})\mathbf{x}_0'$

Después de reemplazar σ^2 por s^2 como en el caso de la regresión lineal simple el intervalo de confianza del $(1 - \alpha)100\%$ para la respuesta media $\mu_{Y/x_{10},x_{20},\ldots,x_{k0}}$ se obtiene a partir de la estadística:

$$T = \frac{\hat{Y} - \mu_{Y/x_{10},x_{20},\ldots,x_{k0}}}{S\sqrt{\mathbf{x}_0'(\mathbf{X'X})\mathbf{x}_0'}}$$

cuya distribución es t con $n-k-1$ grados de libertad.

Luego, el intervalo de confianza del $(1 - \alpha)100\%$ para la respuesta media $\mu_{Y/x_{10},x_{20},\ldots,x_{k0}}$ se obtiene de los siguientes límites de confianza:

$$\hat{y}_0 \mp t_0\, s\sqrt{\mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0'}$$

en donde $t_0 = t_{n-k-1,1-\alpha/2}$. El error estándar es: $s\sqrt{\mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0'}$

Intervalo de predicción para y_0

Para obtener un intervalo de confianza de una respuesta particular y_0 de la variable aleatoria $Y_0 = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i0}$ se considera a la diferencia $\hat{y}_0 - y_0$ como un valor de la variable aleatoria $\hat{Y} - Y_0$ cuya distribución muestral puede demostrase que es normal

con media: $\mu_{\hat{Y}-Y_0} = 0$

varianza: $\sigma^2_{\hat{Y}-Y_0} = \sigma^2[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})\mathbf{x}'_0]$

El intervalo de confianza del $(1 - \alpha)100\%$ para la respuesta media y_0 se determina a partir de la estadística:

$$T = \frac{\hat{Y} - Y_0}{S\sqrt{1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})\mathbf{x}'_0}}$$

cuya distribución es t con $n - k - 1$ grados de libertad.

El intervalo de confianza del $(1 - \alpha)100\%$ para una sola respuesta y_0 se obtiene de los límites de confianza,

$$\hat{y}_0 \mp t_0\, s\sqrt{1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_0}$$

en donde $t_0 = t_{n-k-1,\, 1-\alpha/2}$. El error estándar es: $s\sqrt{1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_0}$

EJEMPLO 13.14

Continuando con el ejemplo 13.12, determine el intervalo de confianza del 95%

a) Para la respuesta media $\mu_{Y/x_1,x_2,x_3}$

b) Para el valor particular y_0, cuando $x_1 = 18$, $x_2 = 17$, $x_3 = 19$.

SOLUCION.

Utilizando el paquete *MCEST*, cuando $x_1 = 18$, $x_2 = 17$, $x_3 = 19$, de la ecuación de regresión se obtiene,

$$y_0 = -2.91069 + 1.01726(18) + 0.04826(17) + 0.1969(19) = 19.962$$

a) El error estándar para la respuesta media es: 1.0833,
Grados de libertad: $n - k - 1 = 10 - 3 - 1 = 6$, $t_{0.975,6} = 2.447$.
Luego, de los limites de confianza:

$$19.962 \mp 2.447 \times 1.0833\,,$$

$$19.962 \mp 2.6508\,.$$

se obtiene el intervalo de confianza,

$$17.31 \le \mu_{Y/18,17,19} \le 22.61.$$

b) El error estándar para la respuesta media es: 1.464
Grados de libertad: $n - k - 1 = 10 - 3 - 1 = 6$, $t_{0.975,6} = 2.447$.

Luego, de los limites de confianza:

$$19.962 \mp 2.447(1.464)\ ,$$
$$19.962 \mp 3.582\ .$$

se obtiene el intervalo de confianza, $16.38 \leq y_0 \leq 23.54$

NOTA. Es evidente que estos intervalos son válidos si la ecuación de regresión muestral es significativa.

13.2.8 Pruebas de hipótesis para los coeficientes de regresión utilizando la distribución de t.

Para probar si los coeficientes β_i de regresión parcial de la población son iguales a cero, se puede establecer la hipótesis nula de dos maneras:

1) La primera prueba es una ***prueba hipótesis global*** de los coeficientes de regresión. Se utiliza para investigar si alguna de las variables independientes tiene un coeficiente de regresión diferente de cero. La hipótesis nula para esta prueba es

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

La hipótesis alternativa es

$$H_1 : \text{Al menos uno de los } \beta_i \neq 0\ .$$

La estadística para esta prueba es $F = \dfrac{CMR}{CME}$ que tiene distribución de probabilidad F con grados de libertad k y $n-k-1$.

Si se acepta H_0, el problema de regresión termina.

Si se rechaza H_0 el problema debe continuar determinando las significación de los coeficientes de regresión por separado.

Esta prueba se ha explicado en la sección 13.2.3.

3) La segunda prueba, es una prueba de coeficientes de regresión **individuales** β_i **por separado**. Se utiliza para determinar cuales de las variables independientes tienen coeficientes de regresión igual a cero y por tanto deberían de eliminarse del modelo de regresión.
Para esta prueba la hipótesis nula para la variable X_i es:

$$H_0 : \beta_i = 0$$

La hipótesis alternativa es: $H_1 : \beta_i \neq 0$ o cualquier alternativa unilateral.

La estadística para esta prueba es :

$$T = \frac{b_i - \beta_i}{\sqrt{c_{ii}\hat{\sigma}^2}} \quad \text{o} \quad T = \frac{b_i - \beta_i}{ES(b_i)}$$

que tiene distribución t-Student con $n - k - 1$ grados de libertad.

$ES(b_i) = \sqrt{c_{ii}\hat{\sigma}^2}$ es el **error estándar** de la estadística b_i.

Si se supone que H_0 es verdadera, entonces, la estadística resultante es:

$$T = \frac{b_i}{ES(b_i)}$$

La decisión de rechazar o aceptar H_0 se puede realizar también a partir de los **intervalos de confianza** para β_i, cuyos limites de confianza son:

$$b_i \mp t_{n-k-1,1-\alpha/2} ES(b_i)$$

EJEMPLO 13.15

Continuando con el ejemplo 13.12 y al nivel de significación 0.05,

a) Realice una prueba de hipótesis global para determinar si algunos de los coeficientes de regresión poblacional no es igual a cero.

b) Si rechaza la hipótesis nula, realice una prueba de hipótesis para coeficientes individuales. ¿Qué variables eliminaría usted?.

c) Halle la ecuación de regresión del modelo reducido y pruebe su significación .

SOLUCION.

a) La hipótesis nula y alternativa para el análisis global son:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

H_1 : al menos una de las β_i es distinto de cero

Los datos del ejemplo 13.12, corridos con el paquete de computo *MCEST* dan la siguiente tabla

ANOVA *para* $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*F* Calculada
Regresión	52.276	3	17.425	17.953
Error	5.824	6	0.971	
Total	58.100	9		

Al nivel de significación 5%, y con los grados de libertad 3 y 6, en la tabla de probabilidades de la F se encuentra el valor crítico $F_{0.95,3,6} = 4.76$.

Dado que el valor calculado, F=17.953 > 4.76, se debería rechazar la hipótesis nula, $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

En consecuencia se puede afirmar que existe regresión lineal poblacional global de Y con X_1, X_2, X_3.

La significación por el método de la probabilidad P (paquete *MCEST*)

$$: \quad P = P[F > 17.953] = 0.002$$

b) Para determinar la significación de los coeficientes de regresión muestral o la contribución al modelo de cada variable independiente X_i se plantean las siguientes hipótesis:

$$H_0 : \beta_i = 0 \text{, contra} \quad H_1 : \beta_i \neq 0 \text{, } i = 1, 2, 3,$$

La estadística de cada prueba es:

$$T = \frac{b_i}{ES(b_i)}$$

cuya distribución es t-Student con $n - k - 1 = 10 - 3 - 1 = 6$ grados de libertad.

A partir de los datos del ejemplo 13.12, se tienen:

Los coeficientes de regresión muestral:

$$b_1 = 1.017, \quad b_2 = 0.0482, \quad b_3 = 0.197$$

Los errores estándares respectivos: $ES(b_1)$=0.179, $ES(b_2)$=0.172, $ES(b_3)$=0.176

Entonces, los valores calculados de la estadística t resultan y las significaciones bilaterales son:

$$\text{Para} \quad \beta_1 = 0 \text{; } t = \frac{b_1}{ES(b_1)} = \frac{1.017}{0.179} = 5.698$$

$$\text{Para} \quad \beta_2 = 0 \text{; } t = \frac{b_2}{ES(b_2)} = \frac{0.0482}{0.172} = 0.281$$

$$\text{Para} \quad \beta_3 = 0 \text{; } t = \frac{b_3}{ES(b_3)} = \frac{0.197}{0.176} = 1.117$$

La **región crítica** para esta prueba bilateral al nivel de significación del 5% es el conjunto:

$$RC = \{T < -2.447 \text{ o } T > 2.447\}$$

Si utilizamos la probabilidad P, las significaciones para b_1, b_2, b_3, son respectivamente: 0.001, 0.788 y 0.307.

Como se puede observar, sólo β_1 es significativo al nivel 5%, en consecuencia sólo X_1 contribuye al modelo de regresión. Las variables X_2, X_3 deberían eliminarse del modelo de regresión.

Por otro lado los **limites de confianza** inferior y superior al 95% son:

Para β_1; $1.017 \mp 2.447 \times 0.179$, $[0.579, 1.455]$

Para β_2; $0.0482 \mp 2.447 \times 0.172$, $[-0.372, 0.468]$

Para β_3; $0.197 \mp 2.447 \times 0.176$, $[-0.234, 0.628]$

Como se puede observar los intervalos de β_2 y β_3 contienen el valor cero. Por lo tanto debemos concluir que $\beta_2=0$ y $\beta_3=0$.

Por el método del intervalo de confianza también, sólo β_1 es significativo.

El paquete *MCEST* resume estos resultados en la tabla que sigue:

Pruebas t de la hipótesis $\beta_1=0$, $\beta_2=0$ y $\beta_3=0$

Modelo	Coeficiente	Error estándar s	*t*	Significación Bilateral	Intervalo de Confianza Inf	Sup
b_0	−2.911					
b_1	1.017	0.179	5.698	0.001	0.579,	1.455
b_2	0.048	0.172	0.281	0.788	−0.372,	0.468
b_3	0.197	0.176	1.117	0.307	−0.234,	0.628

c) La ecuación de regresión lineal muestral resultante de Y con X_1, eliminado las variables X_2, X_3 que no contribuyen al modelo, es:

$$\hat{Y} = -0.455 + 1.117X_1,$$

El error estándar de la pendiente b_1 es 0.147. Grados de libertad =10-2=8.
El intervalo de confianza del 95% para el parámetro β es:

$$1.117 \mp 2.306 \times 0.147, \quad [0.778, 1.456]$$

que no contiene el cero, se concluye que este modelo de regresión es ...ativo.

A este mismo resultado se llega por el método de prueba de hipótesis como se indica la tabla que sigue:

Pruebas t de la hipótesis H_0: $\beta_1 = 0$.

Modelo	Coeficientes	Error estándar	t	Significación Bilateral
b_0	−0.455			
b_1	1.117	0.147	7.619	0.000

13.2.9 Estudio de residuales y violación de supuestos

En el modelo de regresión lineal se supone que los residuos ε_i son independientes y tienen **distribución normal** con **media cero** y **varianza** σ^2.

Los residuales muestrales son las diferencias entre valores observados y los valores estimados o predicciones. Esto es, los residuos son: $e_i = Y_i - \hat{Y}_i$.

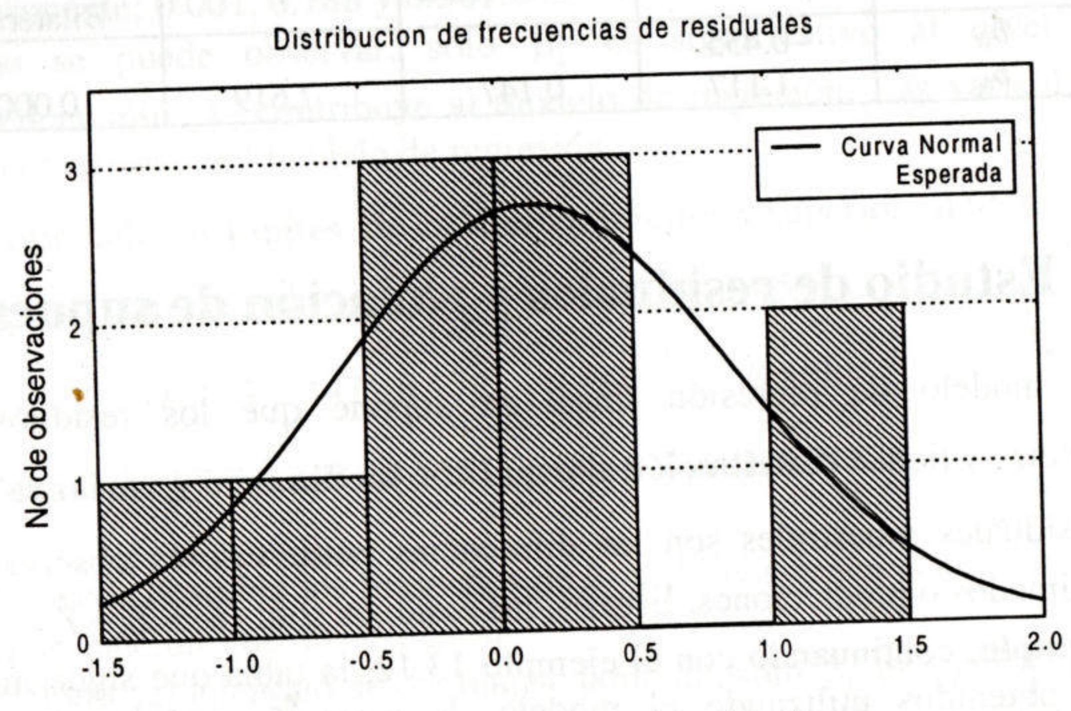
Por **ejemplo,** continuando con el ejemplo 13.12, la tabla que sigue muestra los residuales obtenidos utilizando el modelo de regresión múltiple sin eliminar variables independientes.

Notas finales reales, notas finales predecidas , residuales

Dato Número	Valores Observados	Valores Predecidos	Residuales
1	13.00000	13.32320	.32320
2	15.00000	14.86352	.13648
3	17.00000	17.39225	-.39225
4	16.00000	16.81709	-.81709
5	13.00000	11.86000	1.14000
6	17.00000	16.66461	.33539
7	19.00000	17.58917	1.41083
8	14.00000	15.15697	-1.15697
9	18.00000	17.96358	.03642
10	11.00000	11.36963	-.36963
Minimum	11.00000	11.36963	-1.15697
Maximum	19.00000	17.96358	1.41083
Mean	15.30000	15.30000	.00000
Median	15.50000	15.91079	-.14339

Podemos **verificar el supuesto de normalidad** graficando los residuos muestrales mediante un histograma o un diagrama de tallo y hojas.

Por **ejemplo**, continuando con el ejemplo 13.12 el histograma de residuales, gráfica 132.1, indica que la distribución es algo normal que es uno de los supuestos del modelo de regresión múltiple.



Gráfica 132.1. Histograma de residuales

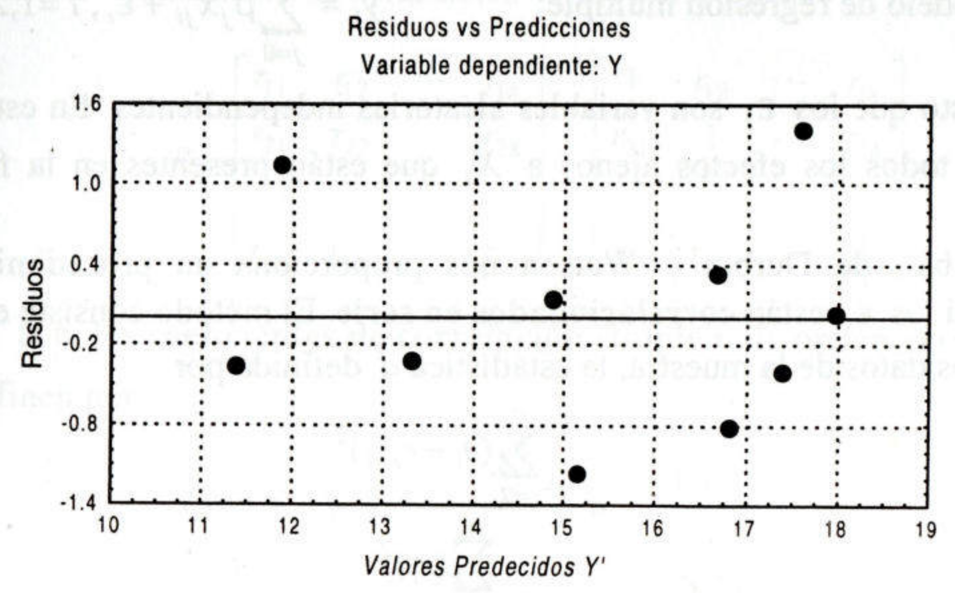
Para **verificar la homocedasticidad** o varianzas constantes (o homogéneas), debemos verificar que los residuales permanezcan constantes para todos los valores de las estimaciones. Para esto, se traza una gráfica de los residuales contra los valores estimados, $\hat{Y}$.

Si no hay mayor variabilidad alrededor de valores grandes de $\hat{Y}$ que alrededor de valores pequeños de $\hat{Y}$ se concluye que se satisface la homogeneidad de varianzas.

Por **ejemplo** la gráfica 132.2 corresponde a la gráfica de residuos versus predicciones del modelo de regresión múltiple **del ejemplo 13.12**, antes de eliminar las variables independientes que no contribuyen al modelo. Esta gráfica revela heterogeneidad de la varianza en una forma ideal

Si la gráfica presenta una forma de cono o embudo, esto es, si a medida que aumenta el valor de $\hat{Y}$ aumenta la variación de los residuales, entonces, se tiene una heterogeneidad de las varianzas o que no se verifica la homocedasticidad.(Ver la figura 132.3)

Si la gráfica de los residuales con las estimaciones respectivas presenta una forma no lineal (parábola, exponencial etc), entonces, se puede concluir que el modelo de regresión está mal especificado (ver figura 132.4)



Grafica 132.2. La gráfica muestra homogeneidad de varianzas

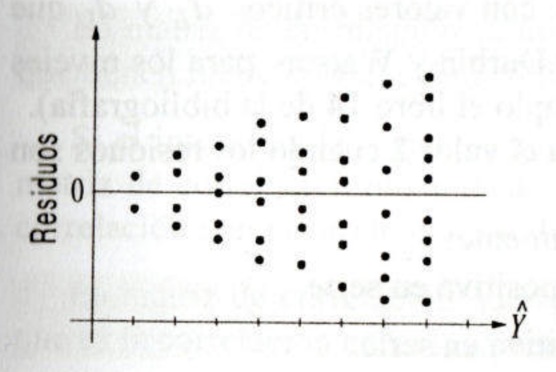


Figura 132.3. La gráfica muestra Heterogeneidad de varianza

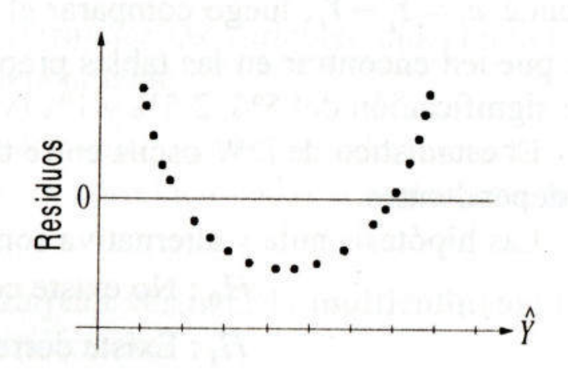


Figura 132.4. La gráfica muestra un modelo que no es lineal

Existen pues tres **tipos de violaciones a los supuestos** del modelo de regresión y que son de fácil detección con el uso de residuales o de gráficas de residuales. Las tres violaciones a los supuestos son las siguientes:

1. Presencia de segregados.
2. Variación heterogénea del error.
3. Modelo mal especificado.

Un *segregado* es un punto en los datos en el que hay un alejamiento de la suposición usual $E(\varepsilon_i) = 0$.

En libros especializados de análisis de regresión y correlación pueden encontrarse el desarrollo de estos temas en forma detallada.

13.2.10 Prueba de Durbin-Watson

En el modelo de regresión múltiple: $y_i = \sum_{j=0}^{k} \beta_j x_{ji} + \varepsilon_i \, , \, i = 1,2,...,n,$

se ha supuesto que los ε_i son variables aleatorias independientes. En esta variable se incluyen todos los efectos ajenos a X_i que están presentes en la función de regresión

La prueba de Durbin y Watson nos proporciona un procedimiento para determinar si los ε_i están **correlacionados en serie**. El método consiste en calcular a partir de los datos de la muestra, la estadística d definida por:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} (e_i)^2}$$

donde $e_i = Y_i - \hat{Y}_i$, luego comparar el valor de d con valores críticos d_L y d_U que se pueden encontrar en las tablas preparadas por Durbin y Watson para los niveles de significación del 5%, 2.5% y 1% (ver por ejemplo el libro 14 de la bibliografía).

El estadístico de DW oscila entre 0 y 4. Toma el valor 2 cuando los residuos son independientes.

Las hipótesis nula y alternativa son respectivamente:

H_0 : No existe correlación positiva en serie

H_1 : Existe correlación positiva en serie.

La regla de decisión es:

Rechazar H_0 si $d < d_L$

No rechazar H_0 si $d > d_U$.

Si $d_L < d < d_U$ la prueba no es concluyente.

Por ejemplo, aplicando a los datos del ejemplo 13.12 la estadística de Durbin-Watson es $d = 2.488$. De la tabla de Durbin-Watson para $n = 15$ observaciones, $k = 2$ variables independientes y un nivel de significación del 5%, se encuentran los valores críticos $d_L = 0.95$ y $d_U = 1.54$.

Dado que $d = 2.5 > d_U = 1.54$, concluimos que no hay autocorrelación en serie o que los residuos son independientes.

13.2.11 Matriz de Correlaciones de orden cero:

La Matriz de correlación de *k* variables $X_1, X_2, \ldots, X_k$ es

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

donde r_{ij} son los **coeficientes de correlación simple o de orden cero de Pearson** que se definen por:

$$r_{ij} = \frac{Cov(X_i, X_j)}{S_{X_i} S_{X_j}} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x})^2} \sqrt{\sum_{k=1}^{n} (x_{jk} - \bar{x}_j)^2}}$$

La matriz de correlación se utiliza para determinar las variables independientes que tienen fuerte correlación con la variable dependiente

Si el interés en la regresión múltiple es usar menos variables independientes, la matriz de correlación nos indica cuales no están correlacionadas o tienen una baja correlación con la variable dependiente.

La matriz de correlación también se utiliza para verificar la **multicolinearidad** que es la correlación entre las variables independientes.

La multicolinearidad puede distorsionar el error estándar de la estimación y nos puede llevar a conclusiones erróneas acerca de cuales variables independientes son estadísticamente significativas.

Por experiencia se sabe que correlaciones entre −0.70 y 0.70 no causan dificultades.

El problema de la multicolinearidad se resuelve omitiendo una de las variables independientes que estén fuertemente correlacionados y volviendo a calcular la ecuación de regresión.

Para probar la significación de los coeficientes de correlación r_{ij} se utiliza la estadística:

$$t = r_{ij} \sqrt{\frac{n-k-2}{1-r_{ij}^2}}$$

que tiene distribución t-Student con $n-k-2$ grados de libertad. Donde k es el número de variables, y n el número de casos.

EJEMPLO 13.16

Continuando con el ejemplo 13.12, halle la matriz de correlación de orden cero (del modelo de regresión múltiple).

a) ¿Cuáles de las variables independientes tiene correlación significativa con la variable dependiente?
b) Analice la multicolinearidad o correlación entre las variables independientes.

SOLUCION.

a) Con el paquete de computo estadístico *MCEST* se obtiene la siguiente matriz de **correlaciones** de Pearson o de orden cero:

Correlaciones de orden cero.

Variable	X_1	X_2	X_3	Y
X_1	**1.000**	**0.052**	**0.478**	**0.937**
X_2	0.052	**1.000**	**–0.226**	**0.044**
X_3	0.478	–0.226	**1.000**	**0.571**
Y	0.937	0.044	0.571	**1.000**

Observe en la última columna que la variable dependiente Y está fuertemente correlacionada sólo con la variable dependiente X_1.

La significación a una cola o unilateral en forma correspondiente se da en la siguiente tabla:

Significación unilateral de los coeficientes de correlación

Variable	X_1	X_2	X_3	Y
X_1	–	0.443	0.081	0.000
X_2		–	0.265	0.452
X_3			–	0.042
Y				–

Observe que, sólo es significativa la correlación de Y con X_1 a un nivel de significación inferior a 4%. Según este criterio deberán eliminarse las variables independientes: X_2, X_3.

b) La correlación simple de X_1 con X_2 es 0.052, de X_1 con X_3 es 0.478 y de X_2 con X_3 es –0.226. Como se observa en la tabla de significación no existe correlación entre las variables independientes. Por lo tanto, no existe multicolionearidad.

13.2.12 Coeficientes de correlación parcial

El **coeficiente de correlación parcial** indica la correlación entre una de las variables independientes y la variable dependiente, manteniendo la otra u otras variables independientes estadísticamente constantes.

El coeficiente de correlación parcial de Y con X_1 cuando X_2 permanece constante se denota por $r_{Y1,2}$ y se define por:

$$r_{y1,2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

donde r_{y1}, r_{y2}, r_{12} son los coeficientes de **correlación simple** respectivamente de Y con X_1, de Y con X_2, y de X_1 con X_2.

Análogamente, el coeficiente de correlación parcial de Y con X_2 cuando X_1 permanece constante es

$$r_{y2,1} = \frac{r_{y2} - r_{y1}r_{21}}{\sqrt{(1-r_{y1}^2)(1-r_{21}^2)}}$$

El coeficiente de correlación parcial de X_1 y X_2 cuando la variable dependiente Y permanece constante es:

$$r_{12,y} = \frac{r_{12} - r_{1y}r_{2y}}{\sqrt{(1-r_{1y}^2)(1-r_{2y}^2)}}$$

El **coeficiente de determinación parcial** es igual al cuadrado del coeficiente de correlación parcial. Este coeficiente se interpreta como el R^2 es decir, es el porcentaje de la varianza de Y explicada por la regresión de Y con la variable dependiente correspondiente, manteniendo constante la otra variable independiente.

EJEMPLO 13.17

Continuando con el ejemplo 13.12, determine

a) Los coeficientes de correlación parcial
b) Los coeficientes de determinación parcial. Realice la interpretación de los mismos.

SOLUCION.

Utilizando el paquete de computo estadístico *MCEST* se obtiene la matriz de correlaciones parciales de las variables: X_1, X_2, X_3, X_4, en donde $X_4 = Y$ que se muestran en la tabla que sigue. (en paréntesis la significación a dos lados o bilateral)

a) El **coeficiente de correlación parcial** de Y con cada X_i manteniendo constantes las demás variables independientes X_s se dan en la columna de la derecha de la tabla (en negritas).

El **coeficiente de correlación parcial** de Y con X_1, manteniendo constantes X_2 y X_3 es $r_{y1,23} = 0.9187$. Solo este coeficiente es significativo.

El **coeficiente de correlación parcial** de Y con X_2, manteniendo constantes X_1 y X_3 es $r_{y2,13} = 0.1141$.

Y el **coeficiente de correlación parcial** de Y con X_3, manteniendo constantes X_1 y X_2 es $r_{y3,12} = 0.41497$.

Correlaciones parciales

Variables	X_1	X_2	X_3	Y
X_1	–	−0.03141 (0.941)	−0.20034 (0.634)	**0.91870** (0.001)
X_2	−0.03141	–	−0.30578 (0.461)	**0.11410** (0.778)
X_3	−0.20034	0.30578	–	**0.41497** (0.307)
Y	0.91870	0.11410	0.04128	–

b) El **coeficiente de determinación parcial** Y con X_1, manteniendo constantes X_2, y X_3 es $R^2_{y1,23} = (0.9187)^2 = 0.844$, entonces, 84.4% es el porcentaje de la varianza de Y que se asocia con X_1 y no con X_2, y X_3.

Del mismo modo se obtiene los siguientes índices de determinación parcial $R^2_{Y2,13} = (0.1141)^2 = 0.013$, y $R^2_{Y3,12} = (0.04128)^2 = 0.0017$.

Como se puede observar del análisis de los coeficientes de determinación múltiple sólo existe ajuste lineal aceptable de Y con X_1.

13.2.13 Modelos de regresión curvilineales

Después de haber estudiado la regresión lineal simple y lineal múltiple, haremos una introducción al estudio de la regresión curvilineal entre dos variables.

El procedimiento usual es inspeccionar el diagrama de dispersión elaborado con los datos muestrales. Si las variables parecen estar relacionadas linealmente, se utiliza un modelo de regresión lineal simple. Cuando las variables no están relacionadas linealmente se debería tratar de convertirla a lineal. Si la transformación lineal no es posible, se debe ajustar a una función matemática curvilínea reconocible según indique el diagrama de dispersión.

A continuación se da una lista de ecuaciones curvilíneas algunas de las cuales pueden ser transformadas a lineal

Ecuación		**Transformación lineal**
Compuesto	$Y = A(B^X)$	$\log Y = \log A + (\log B)X$
Inversa	$Y = a + B/X$	$Y = a + BX'$, con $X' = 1/X$
Logarítmica	$Y = a + b\ln X$	$Y = a + bX'$, con $X' = \ln X$
Potencia	$Y = A(X^B)$	$\log Y = \log A + B\log X$
Crecimiento	$Y = e^{a+bX}$	$\ln Y = a + bX$
Curva S	$Y = e^{a+\frac{b}{X}}$	$\ln Y = a + b/X$
Exponencial	$Y = A(e^{BX})$	$\ln Y = \ln A + BX$
Logística	$Y = 1/(1/u + A(B^X))$	$ln((1/Y-1/u) = ln(A) + ln(B)X$ donde u es un número positivo mayor que el valor máximo de la variable dependiente
Hiperbólica	$Y = 1/(A + BX)$	$Y' = A + BX$, siendo $Y' = 1/Y$
Polinomial	$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX^k$, con $K \ge 2$	

EJEMPLO 13.18

El volumen de ventas mensuales (Y) en miles de dólares y los años de experiencia en ventas (X) registradas de una muestra aleatoria de 6 vendedores de la empresa "COCONA" procesadora de alimentos, se dan en la tabla que sigue.

X	1	2	3	4	5	6
Y	10	40	120	300	800	1500

Se plantean los modelos: $Y = Ae^{BX}$ e $Y = a + bX$,

para relacionar Y **con** X, **¿cuál de los dos modelos se ajusta mejor a los datos?**

SOLUCION.

Si ajustamos a los datos la ecuación no lineal $Y = Ae^{BX}$, su transformación lineal es $\ln Y = \ln A + BX$, esto es

$$Y' = a + BX, \quad \text{donde} \quad Y' = \ln Y, \quad a = \ln A \ .$$

De los datos experimentales se obtiene:

$$\sum X = 21, \quad \sum Y' = 30.481, \quad \sum XY' = 124.16, \quad \sum X^2 = 91, \quad \sum Y'^2 = 172.53$$

$$B = \frac{n\sum XY' - \sum X \sum Y'}{n\sum X^2 - [\sum X]^2} = 0.99876, \qquad a = \overline{y}' - B\overline{x} = 1.58443 ,$$

$$r = 0.9935, \qquad r^2 = 0.987$$

La ecuación lineal de regresión es: $Y' = 1.58443 + 0.99876X$

La ecuación no lineal de regresión es: $Y = anti\ln(Y') = (4.8765)e^{(0.99876)X}$

Si ajustamos la ecuación lineal $Y = a + bX$, de los datos se obtiene:

$$\sum X = 21, \quad \sum Y = 2770, \quad \sum XY = 14650, \quad \sum X^2 = 91, \quad \sum Y^2 = 2996100$$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - [\sum X]^2} = 283.143, \qquad a = \overline{y} - b\overline{x} = -529.33,$$

$$r = 0.904, \qquad r^2 = 0.817$$

La ecuación lineal de regresión es: $Y = -529.33 + 283.143X$

Finalmente comparando los coeficientes de determinación: R^2 o $r^2 = 0.987$, para el modelo no lineal y $r^2 = 0.817$ para el modelo lineal, se concluye que el modelo no lineal se ajusta mejor. La figura 13.3 muestra el ajuste de las curvas.

Figura 13.3. Ajuste lineal y exponencial

El análisis de varianza para ambos modelos se muestra a continuación:

ANOVA: Por el método LINEAL

Fuentes	GL	Suma de cuadrados	Cuadrados medios
Regresión	1	1402972.9	1402972.9
Residuales	4	314310.5	78577.6

$F = 17.85461$. Significación $F = 0.0134$

ANOVA: Por el método EXPONENCIAL

Fuentes	GL	Suma de cuadrados	Cuadrados medios
Regresión	1	17.456691	17.456691
Residuales	4	0.229277	0.057319

$F = 304.55226$. Significación $F = 0.0001$

EJEMPLO 13.19

Se llevó a cabo un experimento para determinar la distancia de frenado a diferentes velocidades de un modelo nuevo de automóvil. se registraron los siguientes datos:

Velocidad X (*K*m./h)	30	40	50	60	70	80
Distancia de frenado Y (metros)	15	25	35	65	100	150

a) Estime la ecuación de regresión cuadrática

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$

b) Estime la distancia de frenado cuando el vehículo viaja a la velocidad de 80 *K*m./h.

c) Determine si el modelo propuesto es el adecuado

SOLUCION

a) Se va a determinar la ecuación muestral :

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$

De los datos, se obtiene las ecuaciones normales:

$$6b_0 + 330b_1 + 19900b_2 = 390$$

$$330b_0 + 19900b_1 + 1287000b_2 = 26100$$

$$19900b_0 + 1287000b_1 + 87550000b_2 = 1825000$$

Las soluciones únicas de este sistema son:

$$b_0 = 65.28434, \quad b_1 = -3.23570, \quad b_2 = 0.05357$$

Luego, el polinomio es:

$$\hat{Y} = 65.28434 - 3.23570X + 0.05357X^2$$

b) Si $X = 90$, entonces $\mu_{Y/80} = 207.99$ metros

c) Utilizando la prueba t para la significación de coeficientes se tiene:

Pruebas t de la hipótesis: β_1=0, β_2=0

Modelo	Coeficientes	Error estándar	t	Significación bilateral
b_0	65.284			
b_1	−3.236	0.516	−6.276	0.011
b_2	0.054	0.005	11.530	0.001

Se concluye, que el modelo es adecuado.

El lector debería resolver los ejemplos 13.17 y 13.18 utilizando un paquete de computo, por ejemplo el *MCEST*.

EJERCICIOS

1. El Gerente de personal de la empresa tabacalera "Puro Peruano" ha estudiado la relación entre las variables: Y = gastos en educación, X_1=salario, y X_2=número de hijos. Una muestra aleatoria de 13 empleados de su empresa ha dado la siguiente tabla de análisis de varianza incompleta:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Valor de Estd. *F*
Regresión	2	C	D	F
Error	A	234	E	
Total	B	14,600		

a) Complete los valores que se muestran con las letras desde la *A* hasta la *E*.
b) Halle el coeficiente de determinación múltiple. De su comentario sobre este valor
c) Plantee las hipótesis para determinar si existe regresión global de la variable Y con respecto a las variables X_1, X_2.
d) Determine la región crítica al nivel de significación 0.05, ¿se puede concluir que existe regresión global de Y con respecto a X_1, X_2?.
e) Calcule la probabilidad P de la prueba de hipótesis
f) Calcule el índice de correlación muestral múltiple. Con este resultado, ¿se puede afirmar que el índice de correlación poblacional es diferente de cero?.
g) Estime la desviación estándar de la regresión poblacional, ¿cómo lo interpreta?.

Rp. a) A=10, B=12, C=14,366, D=7183, E=23.4, F=306.96, b) R^2=0.984, 98% de la varianza de Y se explica por la regresión, c) H_0:β_1=β_2=0, y H_1: al menos uno de los β_1 y β_2 no es cero, d) RC=[4.1,+∞[, se rechaza H_0, si, e) Prob P=0.000, f) R=0.99, F es significativo, entonces, $\rho \neq 0$ en la población. g) $\sigma^2 = CME$ =23.4, σ=4.837, si los errores se distribuyen normalmente el 68% de los residuales son menores que ± σ y el 95 % son menores que ±2σ

2. El gerente de ventas de la compañía de cerveza "Dorada" está estudiando la posibilidad de predecir las ventas semanales (Y) de su producto con base en dos variables predictoras independientes: publicidad (X_1) y transporte (X_2). Se escogieron 13 semanas aleatoriamente y se ha obtenido la siguiente ecuación de regresión muestral (los datos se han obtenido en miles de dólares):

$$\hat{Y} = 3.308 + 0.633X_1 + 1.481X_2$$

a) Interprete los coeficientes de regresión parcial.
b) ¿En cuánto estima la venta de una semana si la compañía gastara $10,000 en publicidad y $1500 en transporte?

c) Un incremento de \$2000 en una semana de la publicidad, ¿en cuánto aumentan las ventas?

d) Un incremento de \$1000 en el transporte de una semana, ¿en cuánto aumentan las ventas?

e) Si además: $SCT = 435.58$, $SCR = 395.12$, calcule el valor de R^2, ¿qué significa este valor con respecto al ajuste del plano de regresión?

f) Calcule el error estándar de la estimación múltiple. ¿Cómo se interpreta?

Rp. b) $\hat{Y} = 11.8595$ o \$11,859.5, c) en promedio 1.266 por mil, d) en promedio 1.48 por mil, e) $R^2 = 0.91$ el 91% de la varianza de las ventas están explicadas por la publicidad y transporte, f) $\sigma^2 = CME = 4.046$, $\sigma = 2.024$, si los errores se distribuyen normalmente el 68% de los residuales son menores que $\pm\sigma$ y el 95 % son menores que $\pm 2\sigma$.

3. El director de marketing de la empresa P&H está estudiando las ventas mensuales de computadoras por departamentos. Se seleccionaron tres variables independientes: Población del departamento (X_1), ingreso percápita en el departamento (X_2) y tasa de desempleo (X_3). La variable dependiente es la venta en \$ (Y). De una muestra aleatoria de 13 departamentos, se obtuvo la siguiente ecuación de regresión muestral múltiple:

$$\hat{Y} = 16000 + 0.35X_1 + 8.24X_2 + 1.02X_3$$

a) Interprete los coeficientes de regresión

b) Si además $SCT = 12452.28$, $SCR = 11242.64$, al nivel de significación 0.05, ¿se puede concluir que existe regresión múltiple global en la población?

c) Halle el coeficiente de determinación, ¿qué porcentaje de la varianza muestral de Y es explicada por la regresión muestral?, ¿es un buen ajuste?

d) Halle el coeficiente de correlación múltiple y determine si es significativo

e) Si los errores estándares de b_1, b_2, b_3, son respectivamente 0.112, 2.132, 1.05, realice la prueba de significación de cada uno de estos coeficientes, ¿qué modelo de regresión recomendaría usted?.

Rp. b) $SCE = 1209.64$, g.l. = 3, 9, $F = 27.8826$, $R.C =]3.86, +\infty[$, se rechaza H_0, c) $R^2 = 0.9028$, d) $R = 0.95$, si es significativo, e) para b_1 es 3.125, para b_2 es 3.865, para b_3 es 0.971, G.L.=9, $T = 2.262$, se descarta la variable X_3.

4. Se ha tomado una muestra de 34 hogares para estudiar sus ahorros: Y (en dólares) con respecto a las variables predictoras: X_1=ingresos, X_2=Número de hijos y X_3=años de convivencia. Los datos muestrales se corrieron con el paquete *MCEST* y se obtuvieron los siguientes resultados

Tabla ANOVA			
Fuente	SC	GL	CM
Regresión	18024.74	3	
Error			
Total	19840.00	33	

Predictoras	Coeficientes	Error estand	Estadist t
Constante	10.0	--	
X_1	0.10	0.025	
X_2	0.12	0.114	
X_3	1.05	0.480	

a) Estime el ahorro de una familia formada hace 6 años, que tiene 3 hijos y cuyo ingreso mensual es \$500.

b) Calcule el coeficiente de determinación, ¿qué porcentaje de la varianza de Y es explicada por la ecuación de regresión muestral?

c) Halle la desviación estándar estimada. ¿Si se verifica la distribución normal de los residuales, ¿qué porcentaje de residuales está comprendidos en $\pm 2\sigma$?.

d) Plantee las hipótesis correspondientes para una prueba global de los coeficientes de regresión poblacional. ¿Cuál es la decisión al nivel de significación 5%?.

e) Si se rechaza la hipótesis nula, realice una prueba de hipótesis de los coeficientes individuales, ¿qué variables se eliminan?.

Rp. a) \$66.66, b) $R^2 = SCR/SCT = 0.9085$, c) s= $(CME)^{1/2}=7.779$, Aprox=0.95, d) SCE=1815.51, gl: 3, 30, 33, F=99.281, Signific=0.000, se rechaza H_0: $\beta_1=\beta_2=\beta_3=0$, e) Estd t para b_1 es 4. para b_2 es 1.05, para b_3 es 2.1875, se elimina la variable X_2.

5. El gerente de ventas de un distribuidor grande de vinos importados realiza un estudio para determinar la relación existente entre, las ventas mensuales (Y en miles de \$), número de distribuidores minoristas (X_1), preferencia (X_2, 0=Nacional, 1=Importado) y número de cajas vendidas (X_3). Los datos de la muestra aleatoria se corrieron con el paquete *MCEST* y se obtuvieron los siguientes resultados:

Tabla ANOVA			
Fuente	SC	GL	CM
Regresión	13838.92	3	
Error	728.36		
Total	14567.28	19	

Predictoras	Coeficientes	Error estand	Estadist t
Constante	5.23	--	
X_1	2.43	0.608	
X_2	0.954	0.681	
X_3	1.05	0.350	

a) ¿Cuál es el tamaño de la muestra?.
b) Calcule el coeficiente de determinación múltiple. Interprete el resultado.
c) Realice una prueba de hipótesis global para determinar si alguno de los coeficientes de regresión poblacional es distinto de cero, al nivel de significación 0.05.
d) Realice una prueba de hipótesis individual para determinar si alguna de sus variables independientes se puede eliminar, al nivel de significación 0.05.
e) Calcule el coeficiente de correlación múltiple, ¿es significativo?.

Rp. a) 20, b) R^2=0.95, c) SCR =728.36, G.L.=3, 16, F =101.33, $R.C$=]3.29, +∞[, se rechaza, H_0:β_1=β_2=β_3=0 , d) para b_1 es 4, para b_2 es 1.4, para b_3 es 3,G.L.=16, T=2.120, se descarta la variable X_2, e) R= 0.975, si es significativo

6. Continuando con el problema 5, de los datos corridos con el paquete *MCEST* se obtiene la siguiente matriz de correlaciones:

Variables	X_1	X_2	X_3	Y
X_1	1	−0.102	0.235	0.96
X_2		1	0.835	0.15
X_3			1	0.90
Y				1

a) ¿Cuál de las variables independientes tiene una correlación más fuerte con la variable dependiente?. Según este criterio, ¿qué variable independiente eliminaría del modelo?
b) ¿Qué variables independientes tienen una correlación más fuerte entres sí?. ¿Parece ser esto un problema?. ¿Por qué?. ¿Cómo se denomina a este problema?.

Rp. a) X_1, X_3 tienen correlación muy fuerte con Y, eliminaría X_2, b) X_1, X_3 tiene alta correlación, este problema se denomina multicolinearidad, puede llevar a conclusiones erróneas acerca de cuáles variables independientes son estadísticamente independientes.

7. En la granja experimental de la Universidad Nacional de Tarapoto se llevó a cabo un experimento para determinar si es posible pronosticar el peso final (Y) de ganado porcino después de 6 meses sobre la base de su peso inicial (X_1) y de la cantidad de alimento que recibe (X_2). El experimento se realizó con una muestra de tamaño 23. Con los datos observados en kilogramos, se obtuvo las siguientes sumas:

$$\sum X_1 = 260, \quad \sum X_2 = 2748, \quad \sum Y = 1339,$$
$$\sum X_1^2 = 3016 \quad \sum X_1 X_2 = 31284, \quad \sum X_1 Y = 15214,$$
$$\sum X_2^2 = 328976, \quad \sum X_2 Y = 160208, \quad \sum Y^2 = 78035$$

a) Escriba las ecuaciones normales del modelo lineal
b) Determine la ecuación de regresión lineal múltiple muestral

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

c) ¿Cuánto es el peso estimado de un cerdo que tuvo un peso inicial de 15kg. y que recibió 100 kg. de alimento?
d) Calcule el error estándar de estimación múltiple $\hat{\sigma}$, ¿qué porcentaje de residuales sería menor que $\pm\hat{\sigma}$
e) Calcule el índice de correlación r o R, el índice de determinación R^2 y el ajustado $\overline{R}^2$. Interprete
f) Determine los coeficientes estandarizados beta.

Rp. b) $\hat{Y} = 26.235 + 0.333X_1 + 0.236X_2$, c) 54.83kg, d) 0.358, 68%, e) R=0.984 R^2=0.968, R^2ajust=0.965, f) $beta_1$=0.323, $beta_2$ =0.665

8. Continuando con el problema 7; suponga que la variable dependiente Y tiene distribución normal.
a) ¿Existe regresión global poblacional al 1%?
b) Si los errores estándares de b_1, b_2 son respectivamente 0.223 y 0.077, realice la prueba de significación de cada uno de estos coeficientes al 5%. ¿Qué variable independiente debería eliminarse del modelo de regresión?
c) Determine la ecuación de regresión lineal del modelo resultante.
d) Pruebe la significación del modelo resultante por el método del intervalo de confianza al 95%.

Rp. a) SCR=79.324, SCE=2.589, SCT=81.913, GL=2,20,22, F=306.376, significación: 0.000, se rechaza $H_0:\beta_1=\beta_2=\beta_3=0$ al 1%. b) t_1=1.496, t_2=3.085, $t(26)$, RC={$T<-2.056$ o $T>2.056$}, luego $\beta_1=0$, se elimina X_1, c) $\hat{Y} = 16.547 + 0.349X_2$, d) ES(b)=0.015, IC de β: [0.319, 0.379]

9. El departamento de producción de una fábrica de confecciones textiles desea explorar la relación entre el número de operarios (X_2) que hacen pantalones, la edad promedio de ellos (X_1), y la cantidad producida por semana (Y). Una muestra aleatoria para realizar el estudio reveló los siguientes datos:

Y	X_1	X_2
30	28	10
45	43	12
52	48	14
55	52	15
70	60	17
75	63	21

a) Determine el sistema de ecuaciones normales.

b) Halle la ecuación de regresión muestral

c) Reduzca el sistema de tres a dos ecuaciones mediante la transformación: $x = X - \overline{X}$

Rp. b) $\hat{Y} = -9.54 + 0.9X_1 + 1.33X_2$,

10. Continuando con el problema 9; suponga normalidad de la variable aleatoria Y Con nivel de significación α=0.05
   a) Pruebe la significación del modelo de regresión múltiple.
   b) Determine la contribución de cada variable predictora al modelo de regresión poblacional.
   c) Obtenga la ecuación de regresión para el modelo reducido y pruebe la significación del modelo lineal resultante por el método del intervalo de confianza.
   d) Halle el error estándar de estimación y calcule los residuales. Ilustre en una gráfica los residuos que son menores y mayores que $\pm \hat{\sigma}_{Y/X}$

Rp. a) SCR=1343.133, SCE=14.367, SCT=1357.5, gl=2,3, F=140.228, significación: 0.001, se rechaza H_0:β_1=β_2=0 , b) error estándar de b_1 es 0.222, de b_2 es 0.728, t_1=4.076, t_2=1.823, significación:0.03 y 0.17, X_2 no contribuye al modelo, c) $\hat{Y} = -8.46 + 1.29X_1$, SCR=1327.225, SCE=30.275, SCT=1357.5, gl=1,4, F=175.357, significación: 0.000, se rechaza H_0: β_1=0.

11. Continuando con el problema 9:
   a) Halle los coeficientes de correlación simple de Pearson o de orden cero de todas las variables que se incluyen en el modelo. ¿Qué variables independientes tiene correlación significativa con la variable dependiente?. ¿Según este criterio que variables independientes se debería eliminar del modelo de regresión
   b) Halle los coeficientes de correlación parcial. Interprete cada uno de ellos.

Rp. a) De Y con X_1 es 0.989, Y con X_2 es 0.965, X_1 con X_2 es 0.938. significación unilateral respectivas: 0.000, 0.001, 0.003 b) De Y con X_1 es 0.92, Y con X_2 es 0.725 es mejor el ajuste con X_1 que con X_2.

12. El gerente de ventas de las llantas "GOOD BY"" que se venden en todo el país realiza un estudio para determinar la relación entre el número de llantas vendidas por mes (Y en cientos) y los gastos en publicidad mensuales (en cientos de dólares), X_1: Radio, X_2: Periódicos. Los datos obtenidos en una muestra de 7 provincias se dan en la tabla que sigue.

Y	X_1	X_2
4	13	15
5	12	16
5	15	18
6	14	17
7	16	17

a) Halle la ecuación de regresión muestral de Y en X_1, X_2

b) Elabore un cuadro de análisis de varianza y determine la validez de la regresión poblacional.

c) Se debería continuar con el problema para determinar las contribuciones separadas de cada una uno de los coeficientes de regresión.

Rp. a) $\hat{Y} = -2.67 + 0.44X_1 + 0.11X_2$, b) SCR=2.531, SCE=2.669, SCT=5.2, gl=2,4, F =0.948, significación: 0.57, se acepta H_0: β_1=0 y β_2=0, NO hay regresión global. c) NO ninguna de las dos variables independientes contribuye a la regresión.

13. Una compañía grande maneja una cadena de tiendas al menudeo. Como una forma de medir la eficiencia de las distintas tiendas, la gerencia respectiva, estudia la relación existente entre:

Y: Ventas promedios diarios en cientos de dólares
X_1: Número de empleados por tienda.
X_2: Tamaño de cada tienda por metros cuadrados.
X_3: Ubicación: Lima=1, Provincias=0
n = 13 número de tiendas.

Si se tiene la siguiente información:

$$\sum X_1 = 84, \quad \sum X_2 = 272, \quad \sum X_3 = 8, \quad \sum Y = 77,$$
$$\sum X_1^2 = 564, \quad \sum X_1X_2 = 1791, \quad \sum X_1X_3 = 52, \quad \sum X_1Y = 516,$$
$$\sum X_2^2 = 5750, \quad \sum X_2X_3 = 169 \quad \sum X_2Y = 1643,$$
$$\sum X_3^2 = 8, \quad \sum X_3Y = 48, \quad \sum Y^2 = 475$$

Suponga que el modelo de regresión que se plantea es :

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

a) Halle las ecuaciones normales del modelo.

b) Determine la ecuación de regresión muestral por el método de mínimos cuadrados.

c) Elabore un cuadro de análisis de varianza y determine la validez de la regresión poblacional.

d) Determine r o R, R^2 y R^2 ajustado. Interprete

e) Calcule el error estándar de la estimación

f) Si los errores estándares de b_1, b_2, b_3, son respectivamente 0.288, 0.174, 0.247, realice la prueba de significación de cada uno de estos coeficientes al nivel de significación de 0.05, ¿qué modelo de regresión lineal recomendaría usted?.

g) Obtenga la ecuación de regresión muestral eliminando las variables independientes que no contribuyen al modelo.

Rp. b) $\hat{Y} = -4.688 + 0.133X_1 + 0.468X_2 - 0.0587X_3$, c) SCR=17.356, SCE=1.567, SCT=18.923, GL=3,9,12, F=33.224, significación: 0.000, se rechaza H_0:$\beta_1=\beta_2=\beta_3=0$, d) R=0.958, R^2=0.917, R^2ajust=0.889, e) 0.4174, f) GL=9, Tcal: 0.463, 2.691 y −0.239 signific: 0.654, 0.025, 0.817. Se eliminan X_1 y X_3

14. Se ha tomado una muestra aleatoria para estudiar la relación entre las siguientes variables, Y = Salarios ($), X_1=Edad, X_2=Años de servicio, X_3=Número de hijos. Los datos experimentales son:

Y	X_1	X_2	X_3
600	33	5	6
620	34	12	3
500	35	8	4
700	34	4	2
800	35	7	3
850	40	9	0
750	38	4	2
900	29	6	3
500	39	7	4

a) Halle la ecuación de regresión lineal múltiple muestral
b) Determine los coeficientes estandarizados beta. Interprete
c) Efectúe una prueba de hipótesis global para determinar si alguno de los coeficientes de regresión es diferente de cero, al nivel de significación α=0.05.
d) Se pueden realizar predicciones válidas.

Rp. a) $\hat{Y} = 1840.595 - 23.691X_1 - 11.892X_2 - 77.708X_3$, b) $beta_1$=−0.549, $beta_2$ =−0.209, $beta_3$=−0.882, c) SCR=117712.12, SCE=52976.764, SCT=170688.89, gl=3,5,8, F=3.703, significación: 0.096, se acepta H_0: $\beta_1=\beta_2=\beta_3=0$. NO existe regresión múltiple. d) NO, el lector debería verificar que tampoco hay regresión doble ni simple.

15. Se han seleccionado en forma aleatoria a diez fábricas de confecciones textiles para evaluar los efectos del nivel de producción y del índice de costos de mano de obra y materia prima en los costos de manufacturas Las variables son:

Y: Costo promedio de manufactura en dólares
X_1: Nivel de producción como porcentaje de la capacidad fijada
X_2: Indice en porcentajes de los costos de mano de obra y materia prima.

Datos recopilados durante los últimos 10 trimestres.

Trimestre	Y	X_1	X_2
1	3.62	84	82
2	4.21	79	95
3	4.30	80	108
4	5.45	70	115
5	6.63	55	135
6	5.70	60	125
7	5.05	72	114
8	4.00	92	96
9	4.35	96	98
10	4.24	100	110

a) Halle la ecuación de regresión muestral de Y en X_1, X_2

b) Diga, qué tan bien explican o predicen esos factores al costo promedio de manufactura.

c) Para el siguiente trimestre, se espera que los costos de mano de obra y materia prima suban a 145%, y se espera que el nivel de producción suba al 80% de la capacidad, ¿qué costo de manufactura esperaría usted.?.

d) Realice una prueba de hipótesis global de los coeficientes de regresión. ¿Es alguno de ellos diferente de cero al nivel de significación de 0.05?

e) Realice una prueba de hipótesis para determinar si la contribución al modelo lineal de cada variable independiente es significativa.

f) Trace un histograma de los residuales, ¿le parece razonable el supuesto de normalidad para los errores o residuales?

g) Trace una gráfica de los residuales y de los valores estimados, ¿ve usted alguna violación del supuesto de homocedasticidad?

Rp. a) $\hat{Y} = 2.19 - 0.0226X_1 + 0.04X_2$, b) R^2=0.94 el 94% de la varianza de los costos se explica por el nivel de producción y el costo de la mano de obra , c) Y_{Estim}=6.182 d) *SCR*=7.28, *SCE*=0.49, *SCT*=7.77, gl=2,7,9, F =51.938, significación: 0.000, se rechaza H_0, e) error estándar de b_1 es 0.0085, de b_2 es 0.008, t_1=−2.67, t_2=4.998, significación: 0.03 y 0.002, ambas variables independientes contribuyen significativamente en el modelo de regresión.

16. Se han seleccionado en forma aleatoria a diez sucursales de una compañía de ventas al por menor para evaluar los efectos de la población y del ingreso en las ventas en cada distrito de ventas. Las variables son
Y : Venta en miles de dólares,
X_1: Población en millares,
X_2: Ingreso en millones de dólares.

Y	X_1	X_2
4	2	5
5	3	6
5	4	6
6	5	7
7	6	8

7	7	9
8	8	10
9	9	11
10	10	13
11	11	15

a) Analice la tendencia de Y con X_1 y de Y con X_2 utilizando diagrama de dispersión en cada caso
b) Determine las ecuaciones normales del modelo de regresión lineal muestral
c) Halle la ecuación de regresión muestral de ventas sobre población e ingresos utilizando el método de mínimos cuadrados
d) Interprete los coeficientes de regresión parcial utilizando coeficientes beta.
e) Realice una prueba de hipótesis global al nivel de significación 0.05. ¿Son algunos de los coeficientes de regresión distintos de cero?. Calcule la probabilidad P.
f) Realice una prueba de hipótesis individual de los coeficientes de regresión. ¿Se debería eliminar alguna de las variables independientes al nivel de significación 0.05?, ¿ y al nivel 0.01?. Calcule la probabilidad P.
g) ¿Cuánto es el coeficiente de determinación múltiple?. Interprete este cvalor. ¿Por qué se debe calcular siempre el coeficiente de determinación múltiple ajustado?
h) Halle el coeficiente de regresión múltiple. ¿Es significativo al 5%?

Rp. c) $\hat{Y} = 1.499 + 0.35X_1 + 0.38X_2$, d) beta$_1$=0.461, beta$_2$=0.54, e) SCR=47.209, SCE=0.391, SCT=47.6, gl=2,7,9, $F = 423.025$, Prob P=0.000, se rechaza H_0:$\beta_1=\beta_2=0$, f) error estándar de b_1 es 0.124, de b_2 es 0.115, t_1=2.837, t_2=3.321, significación: 0.025 y 0.013, X_1 y X_2 contribuyen al modelo al nivel 0.05, g) R^2=0.99 el 99% de la varianza de las ventas es explicada por los ingresos de la población, h) 0.996 Si por el ANOVA.

17. Continuando con el problema 16
a) Halle la matriz de correlación de orden cero (o simple de Pearson). ¿Qué variables independientes tiene correlación significativa con la variable dependiente?. Según este criterio, ¿qué variables independientes se debería eliminar del modelo de regresión?
b) Analice la multicolinearidad (o correlación entre variables independientes) ¿Qué variables independientes deberíamos omitir del modelo?. ¿Por qué?
c) Halle la matriz de correlación parcial. Realice un comentario de la correlación parcial de Y con X_1 y de Y con X_2. ¿Qué variables independientes se debería eliminar del modelo de regresión?

Rp. a) De Y con X_1 es 0.989, Y con X_2 es 0.992, X_1 con X_2 es 0.978. significación unilateral respectivas: 0.000, 0.001, 0.003 b) De Y con X_1 es 0.731, Y con X_2 es 0.782 es mejor el ajuste con X_2 .

18. Continuando con el problema 16
a) Calcule los residuales de la ecuación de regresión muestral resultante

b) Calcule el error estándar de estimación $\hat{\sigma}$. ¿Qué porcentaje de los residuales de la muestra son menores de $\pm\hat{\sigma}$?

c) Desarrolle un histograma con los residuales. ¿Es razonable concluir que se satisface el supuesto de normalidad?

d) Grafique los residuales contra los valores estimados $\hat{Y}$. ¿Observa usted que se viola el supuesto de homocedasticidad?

Rp. a) Estimad: 4.10, 4.83, 5.18, 5.91, 6.64, 7.36, 8.11, 8.84, 9.95, 11.06. Residuos: −0.10, 0.17, −0.18, 0.09, 0.36, −0.38, −0.11, 0.16, 0.05, −0.06. b) 0.236, 20%,

19. Se obtuvieron los siguientes datos estadísticos de 14 departamentos del país para realizar un estudio de la maternidad adolescente (Y en %), familia en extrema piobreza (X_1 en %) y familia desintegrada (X_2)

Dpto	Y	X_1	X_2
1	12	20	15
2	14	25	15
3	16	22	20
4	18	18	24
5	20	15	26
6	16	14	19
7	10	19	17

Dpto	Y	X_1	X_2
8	13	14	12
9	22	28	33
10	18	24	18
11	11	15	18
12	15	19	28
13	13	14	18
14	29	5	22

a) Utilice el método de matrices para determinar las ecuaciones normales: $\mathbf{(X'X)b = X'Y}$

b) Halle el vector solución $\mathbf{b = (X'X)^{-1}X'Y}$ y escriba la ecuación de regresión muestral

c) Pruebe la hipótesis nula de que no existe regresión global en la población al nivel de significación 5%. ¿Cuál es la probabilidad P de esta prueba?

d) Analice la ***matriz de correlación*** simple (o de orden cero) para determinar qué variable independiente tiene fuerte o débil correlación con la variable dependiente. ¿Según este criterio qué variable independiente debería eliminar del modelo?

e) ¿Existe multicolinearidad entre las dos variables independientes?

f) Analice la ***matriz de correlación*** parcial. Realice un comentario de la correlación parcial de Y con X_1 y de Y con X_2. ¿Qué variable independiente se debería eliminar del modelo de regresión?

g) Utilice la técnica del intervalo de confianza para determinar si alguna de las variables independientes X_1 y X_2 no contribuyen al modelo de regresión múltiple al nivel de confianza de 0.95.

h) Halle la ecuación de regresión muestral del modelo resultante. ¿Existe regresión simple en la población al nivel de significación 1%?. ¿Qué porcentaje de la varianza de Y se explica por la regresión?

Rp. a) $14b_0+252b_1+285b_2=227$, $252b_0+4978b_1+5214b_2=3987$, $285b_0+5214b_1+6225b_2=4832$, b) $b_0=10.692$, $b_1=-0.331$, $b_2=0.564$, c) $SCR=151.78$, $SCE=176.578$, $SCT=328.357$, GL=2,11,13, $F=4.728$, significación:0.033, se rechaza H_0: $\beta_1=\beta_2=0$, d) De Y con X_1 es -0.26, Y con X_2 es 0.566, signific unilateral: 0.185, 0.017, e) De X_1 con X_2 es 0.194 signific 0.253, f) De Y con X_1 es -0.457, de Y con X_2 es 0.651, es mejor el ajuste con X_2. g) Error estándar de b_1 es 0.194, de b_2 es 0.199, IC al 95%: [−0.759, 0.096], [0.127, 1.001], se elimina X_1, h) $Y'=6.068+0.498X_2$, signif. 0.035, No, 0.32.

20. El gerente de recursos humanos de una empresa grande que tiene más de 800 empleados realiza un estudio de los salarios de los empleados utilizando una muestra aleatoria de tamaño 24. De cada empleado recabó

X_1: Edad

X_2: Años de servicio

X_3: Género: Hombre =1, Mujer =0

Y: Salario mensual en dólares

Los datos obtenidos son los siguientes:

Empleado	X_1	X_2	X_3	Y
1	20	0.5	1	50
2	20	1	0	80
3	21	1	0	90
4	23	3	1	100
5	24	5	1	120
6	25	6	1	150
7	26	7	1	160
8	26	7	1	180
9	26	7	0	190
10	26	8	0	195
11	3	9	1	200
12	31	10	1	250

Empleado	X_1	X_2	X_3	Y
13	35	12	1	280
14	36	15	0	300
15	37	16	1	320
16	38	16	1	350
17	39	17	1	390
18	40	18	0	420
19	48	19	1	480
20	50	23	0	430
21	52	24	0	490
22	56	26	1	510
23	62	30	1	550
24	64	32	1	590

a) Determine la ecuación de regresión muestral utilizando la variable salario como variable dependiente.

b) Determine el valor del coeficiente de determinación múltiple. De su comentario sobre la bondad del ajuste

c) Desarrolle una prueba de hipótesis global para determinar si alguno de los coeficientes de regresión es diferente de cero.

d) Desarrolle una prueba de hipótesis individual utilizando el método de intervalos de confianza para determinar si se puede eliminar alguna de las variables independientes.

Rp. a) $\hat{Y}=28.614+1.945X_1+14.755X_2-5.835X_3$, b) $R^2=0.793$, R^2ajustado=0.737, c) $SCR=599587.32$, $SCE=17436.643$, $SCT=617023.96$, gl=3,20,23 $F=229.244$, signific 0.000, se rechaza H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$, d) Error estándar de b_1 es3.394, de b_2 es 4.88, de b_3 es 12.837, IC al 95%: [−5.135, 9.025], [4.576, 24.934], [−32.613, 20.943] contribuye X_2.

21. Continuando con el problema 20
   a) Halle la matriz de correlación de orden cero (o simple de Pearson). ¿Qué variables independientes tiene correlación significativa con la variable dependiente?. Según este criterio, ¿qué variables independientes se debería eliminar del modelo de regresión?
   b) Analice la multicolinearidad (o correlación entre variables independientes) ¿Qué variables independientes deberíamos omitir del modelo?. ¿Por qué?
   c) Halle la matriz de correlación parcial. Realice un comentario de la correlación parcial de Y con X_1 y de Y con X_2, Y con X_3. ¿Qué variables independientes se debería eliminar del modelo de regresión?

Rp. a) De Y con X_1 es 0.979, Y con X_2 es 0.985, Y con X_3 es 0.053. signific unilateral respect: 0.000, 0.000, 0.402 b) De X_1 con X_2 es 0.991, X_1 con X_3 es 0.077, X_2 con X_3 es 0.070. signific unilateral respect: 0.000, 0.360, 0.372, c) De Y con X_1 es 0.127, Y con X_2 es 0.56, Y con X_3 es –0.101 es mejor el ajuste con X_2.

22. Continuando con el problema 20,
   a) Si elimina alguna de las variables independientes, determine nuevamente la ecuación de regresión muestral. Analice nuevamente si se puede eliminar alguna de las variables independientes.
   b) Calcule el error estándar de estimación $\hat{\sigma}$, y los residuales de la ecuación de regresión muestral resultante. Dibuje un diagrama de dispersión para indicar los residuales menores y mayores de $\pm\hat{\sigma}$.
   c) Utilice un histograma para determinar si se viola el supuesto de normalidad en el modelo de regresión
   d) Grafique los residuales con respecto a los valores ajustados $\hat{Y}$. ¿Indica el diagrama que se viola el supuesto de homocedasticidad?.

Rp. a) $\hat{Y}$ = 58.527+17.505X_2, SCR=599144.91, SCE=17879.05, SCT=617023.96, gl=1,22,23, F=737.242, signific: 0.000, se rechaza H_0: β_1=0, b) $\hat{\sigma}$ =28.5076.

23. El gerente de personal de una empresa textil de Gamarra utilizó a 30 operarios en un estudio para determinar la relación entre las siguientes variables:

Y: Comportamiento hacia el trabajo (Prueba calificada de 0 a 20)
X_1: Horas semanales de trabajo
X_2: Servicios en el hogar: Teléfono, TVCable, Intenet (0=Uno de los tres, 1=Dos de los tres, 2=Los tres)
X_3: N# de prendas que confecciona por semana
X_4: Años de experiencia

Y	X_1	X_2	X_3	X_4
5	50	0	30	0.6
5	53	0	31	1.0

Y	X_1	X_2	X_3	X_4
14	70	1	38	8.0
14	70	1	39	8.4

6	55	0	31	1.5	15	72	1	39	8.6
6	58	1	32	1.8	15	72	0	40	8.9
8	61	1	32	2.0	16	73	0	41	9.0
9	62	0	33	2.4	16	74	0	42	9.0
9	62	2	34	2.8	16	74	1	43	9.1
10	63	0	35	3.0	16	75	0	44	9.2
10	63	1	35	3.5	17	75	0	44	9.8
10	65	2	36	4.0	17	76	1	45	10.0
10	65	0	36	4.6	17	77	0	45	10.2
10	69	1	36	5.0	18	78	1	46	10.8
11	68	0	37	5.8	18	78	1	47	11.0
12	69	1	37	6.0	19	79	1	48	11.5
13	69	1	38	6.7	20	80	2	49	11.6

a) Halle la ecuación de regresión muestral: $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$

b) Determine el valor del coeficiente de determinación múltiple. De su comentario sobre la bondad del ajuste

c) Desarrolle una prueba de hipótesis global para determinar si alguno de los coeficientes de regresión poblacional es diferente de cero. Utilice el método de la probabilidad P.

d) Calcule el coeficiente de correlación múltiple. ¿Es significativo este coeficiente al nivel de significación 0.01?

Rp. a) $\hat{Y} = -8.163 + 0.138X_1$ +0..02774X_2 + $0.187X_3 + 0.639X_4$, b) R^2=0.985,R^2ajustado=0.982, c) *SCR*=551.326, *SCE*=8.541, *SCT*=559.867, gl=4,25,29, *F* =403..434, Prob. *P*=0.000, se rechaza H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$, d) 0.992, Si,.

24. Continuando con el problema 23

a) Halle la matriz de correlación de orden cero (o simple de Pearson). ¿Qué variables independientes tiene correlación significativa con la variable dependiente?. Según este criterio, ¿qué variables independientes se debería eliminar del modelo de regresión?

b) Analice la multicolinearidad (o correlación entre variables independientes) ¿Qué variables independientes deberíamos omitir del modelo?. ¿Por qué?

c) Halle la matriz de correlación parcial. Realice un comentario de la correlación parcial de Y con X_1 y de Y con X_2, Y con X_3, Y con X_4. ¿Qué variables independientes se debería eliminar del modelo de regresión?

Rp. a) De Y con X_1 es 0.977, Y con X_2 es 0.158, Y con X_3 es 0.975, Y con X_4 es 0.987, significac. unilat: 0.000, 0.202, 0.000, 0.000 , b) X_1 con X_2 es 0.196, X_1 con X_3 es 0.957, X_1 con X_4 es 0.97, X_2 con X_3 es 0.167, X_2 con X_4 es 0.127, X_3 con X_4 es 0.968, signif unil : 0.149, 0.000, 0.000, 0.189, 0.251, 0.000 c) De Y con X_1 es 0.413, Y con X_2 es 0.032, Y con X_3 es 0.413, Y con X_4 es 0.630.

25. Continuando con el problema 23,
   a) Desarrolle una prueba de hipótesis individual utilizando el nivel de significación 0.05 para determinar si se debe eliminar alguna de las variables independientes.
   b) Si elimina alguna de las variables determine nuevamente la ecuación de regresión muestral. Analice nuevamente si alguna de las variables independientes se eliminan.
   c) Determine los residuales de la ecuación de regresión muestral resultante. Utilice un histograma para determinar si se viola el supuesto de normalidad en el modelo de regresión
   d) Grafique los residuales con respecto a los valores ajustados, $\hat{Y}$. ¿Indica el diagrama que se viola el supuesto de homocedasticidad?.

Rp. a) Error estándar de b_1 de 0.061, de b_2 es 0.174, de b_3 es 0.082, de b_4 es 0.157, GL=25, t_1=2.227, t_2=0.159, t_3=2.27, t_4=4.057, significación: 0.032, 0.875, 0.032 y 0.000 se elimina X_2. b) $\hat{Y} = -8.323 + 0.141X_1 + 0.188X_3 + 0.632X_4$ SCR=551.317, SCE=8.55, SCT=559.867 gl=3,26,29, F=558..854, significación: 0.000, se rechaza H_0: $\beta_1=\beta_3=\beta_4=0$. Prueba t: Error estándar de b_1 de 0.256, de b_3 es 0.236, de b_4 es 0.511, GL=26, t_1=2.441, t_3=2.34, t_4=4.25, significación: 0.022, 0.027, y 0.000.

25*. Con los datos de la hoja de cálculo del *estudio socioeconómico de los estudiantes universitarios de Lima* (ver apéndice) realice una análisis de regresión múltiple eligiendo la variable X_8 como variable dependiente.

Regresión curvilínea

1. Los siguientes datos son los precios de venta en dólares Y de una marca de automóviles usados X años:

X	1	2	3	4	5	6
Y	6350	5695	5790	–	4985	4890

   a) Ajuste a los datos la **curva de crecimiento:** $Y = e^{b_0 + b_1 X}$.
   b) Estime el precio de venta de un automóvil que tenga 4 años de uso
   c) Analice la bondad del ajuste con el coeficiente de determinación.
   d) Pruebe la significación de la ecuación muestral mediante un ANOVA

Rp. a) $Y = e^{8.7887 - 0.051X}$, r=–0.9659, R^2=0.933 b) $\hat{y} = 5349.2$, b) SCR=0.04473, SCE=0.03216, GL=1, 3, F=41.72, significación: 0.0075

2. Los siguientes datos son la inversión neta (*Y*) y la tasa de interés (*X*)

X	2.5	3	4	5	5.5	6	7
Y	12.5	10	7	4.5	4	3	3.5

Se plantean dos modelos para relacionar *Y* con *X*

Potencia: $Y = AX^B$ y Lineal: $Y = a + bX$

a) ¿Qué modelo se ajusta mejor a los datos?, ¿por qué?
b) Pruebe la significación de la ecuación que mejor se ajusta

Rp. Potencia, $Y = 46.497(X)^{-1.42}$, $R^2=0.956$. Lineal, $Y=16.28-2.1X$, $R^2=0.884$. Potencia ajusta mejor. b) ANOVA Potencia: $SCR=1.7224$, $SCE=0.0796$, GL=1, 5, $F=108.18$, signific: 0.000.

3. Dados los siguientes datos *X*: Ingreso, *Y*: Consumo en miles de dólares

X	4	8	12	16	20	24	28	32
Y	3	7	10	14	15	20	21	24

Se plantean dos modelos curvilíneos para relacionar *Y* con *X*

Compuesta: $Y = b_0(b_1^X)$ y Cuadrática: $Y = b_0 + b_1X + b_2X^2$

a) ¿Qué modelo se ajusta mejor a los datos?, ¿por qué?
b) Pruebe la significación de la ecuación que mejor se ajusta

Rp. Compuest: $Y=3.669(1.0683)^X$, $R^2=0.869$. Cuadrátic $Y=-0.6429+0.9792X-0.00675X^2$, $R^2=0.991$. b) ANOVA Cuadrática: $SCR=368.024$, $SCE=3.476$, GL=2, 5, $F=264.67$, signific: 0.000.

4. La presión *Y* (kg./cm^2.) de un gas correspondiente a diferentes volúmenes *X* (cm^3.) se registró en la siguiente tabla:

X	50	60	70	80	90	100
Y	79.7	65.1	52.6	36.8	25.7	18.7

Se plantean dos modelos para relacionar *Y* con *X*

Exponencial: $Y = b_0(e^{b_1X})$ y Curva *S*: $Y = e^{b_0 + \frac{b_1}{X}}$

a) ¿Qué modelo se ajusta mejor a los datos?, ¿por qué?
b) Pruebe la significación de la ecuación que mejor se ajusta

Rp. Exp: $Y=381.296(e^{-0.0297X})$, $r^2=0.987$. Curva S, $b_0=1.707$, $b_1=142.574$, $r^2=0.906$. Exp. ajusta mejor. b) ANOVA Exp.: $SCR=1.5434$, $SCE=0.0203$, GL=1, 4, $F=303.61$, signif: 0.000.

5. Los siguientes datos son X: Precios en soles, Y: Cantidades vendidas (en miles de unidades) de un artículo en el mercado en un periodo de 8 meses.

X	4	8	12	14	18	23	28	32
Y	240	200	150	130	100	80	60	30

Se plantean dos modelos para relacionar Y con X

Logarítmica: $Y = b_0 + b_1 \ln(X)$ y Inversa: $Y = b_0 + (b_1 / X)$

a) ¿Qué modelo se ajusta mejor a los datos?, ¿por qué?

b) Pruebe la significación de la ecuación que mejor se ajusta

Rp. Logarítmica: Y=395.297−101.45(ln$X$), R^2=0.98, Inversa: Y= 46.42+(889.117/X), R^2=0.824.
b) ANOVA de Log: SCR=34667.89, SCE=719.62, GL=1, 6, F=289.05, signif: 0.000.

6. Las distancias de parada Y (en metros) de un automóvil que viaja a una velocidad X (en k/h) en el instante en que se observa el peligro se dan en la siguiente tabla:

X	20	30	40	60	80	90	100	110	120
Y	8	10	20	30	50	70	90	150	250

a) Ajuste a los datos la **curva cúbica:** $Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3$

b) Analice la bondad del ajuste con el coeficiente de determinación.

c) Pruebe la significación de la ecuación muestral mediante un ANOVA

Rp. a) Y= −76.87+5.83X −0.11X^2+0.0007X^3, b) R^2=0.989, c) ANOVA: SCR=50335.489, SCE=552.511, GL=3, 5, F=151.84, significación: 0.000.

7. El volumen mensual de ventas (Y) en miles de dólares y sus años de experiencia en ventas (X) de 8 vendedores profesionales de una compañía se dan en la siguiente tabla:

X	0	1	2	3	4	5	6	7
Y	20	30	45	80	120	180	250	300

a) Ajuste a los datos la **curva logística:** $Y = 1/((1/u) + b_0 (b_1^X))$ con u=310.

b) Analice la bondad del ajuste con el coeficiente de determinación.

c) Pruebe la significación de la ecuación muestral mediante un ANOVA

Rp. a) b_0=0.0793, b_1=0.4465, R^2=0.936. b) ANOVA: SCR=27.3, SCE=1.9, GL=1, 6, F=87.89, significación: 0.0001

Capítulo 14

INTRODUCCION AL ANALISIS DE LAS SERIES DE TIEMPO.

14.1 Introducción

Las serie de tiempo o serie cronológica es un conjunto de datos observados en forma secuencial, generalmente en intervalos de tiempo iguales.

Son ejemplos de series de tiempo:

- Cotización diaria del dólar.
- Ventas mensuales de un determinado producto.
- Producción anual de una fábrica.
- Número de trabajadores que laboran en una compañía durante varios años
- Etc.

Diversos tipos de cambios y movimientos ocurren en una serie de tiempo. Estos cambios son causados por factores o componentes que afectan a la serie algunos a largo plazo y otros a corto plazo.

Existe una amplia gama de aplicaciones donde se realizan análisis de series de tiempo. Este es un tema muy avanzado y especializado. Algunos métodos de análisis incluyen técnicas muy elaboradas de las que no pueden incluirse en este texto básico.

El objetivo de este capítulo es presentar en forma introductoria el análisis de las series de tiempo, como determinar la ecuación de tendencia lineal y la medición de los tipos de cambios o movimientos que influyen en la serie a través del tiempo.

Definición. Una **serie de tiempo** es un conjunto de valores,

$$y_1, y_2, ..., y_k$$

de una variable aleatoria Y observados secuencialmente en los periodos de tiempo (iguales) $t_1, t_2, ..., t_k$ (años, meses, trimestres etc.).

Gráfica.

La serie de tiempo se representa mediante una **gráfica de líneas**. En el eje vertical se representan los valores de la serie, mientras que en el eje horizontal se representan los periodos de tiempo.

En la gráfica de serie de tiempo se destacan, entre otros aspectos, los *picos* y los *valles*. Un ***pico*** se produce en un punto donde la tendencia creciente cambia a tendencia decreciente. Un ***valle*** se produce cuando la tendencia decreciente cambia a tendencia creciente.

En general un gráfico de serie de tiempo puede considerarse como el trazo que produce un punto que se mueve a través del tiempo impulsado por una combinación de fuerzas que pueden ser económicas, sociales, psicológicas, etc.

EJEMPLO 14.1.

Las producciones anuales (en millones de unidades) durante 12 años de una compañía ficticia se dan en el cuadro 14.1. Grafique esta serie de tiempo en un sistema cartesiano.

Cuadro 14.1. *Producción de 1987 a 1998 (en millones de unidades)*

Año	X	Producción: Y	año	X	Producción: Y
1987	0	2	1993	6	13
1988	1	3	1994	7	10
1989	2	5	1995	8	17
1990	3	9	1996	9	14
1991	4	12	1997	10	22
1992	5	16	1998	11	24

SOLUCION.

Sea X la variable con que se representa los años codificados, esto es, $X = 0$ representa al año 1987, $X = 1$, representa al año 1988, etc.

Además, representemos por Y las producciones anuales en millones de unidades.

La gráfica de la serie de tiempo es la figura 14.1

Observar que se produce una cumbre o pico en el año 1992 ($X = 5$) y otro en 1995 ($X = 8$), un valle en 1994 ($X = 7$) y otro en 1996 ($X = 9$).

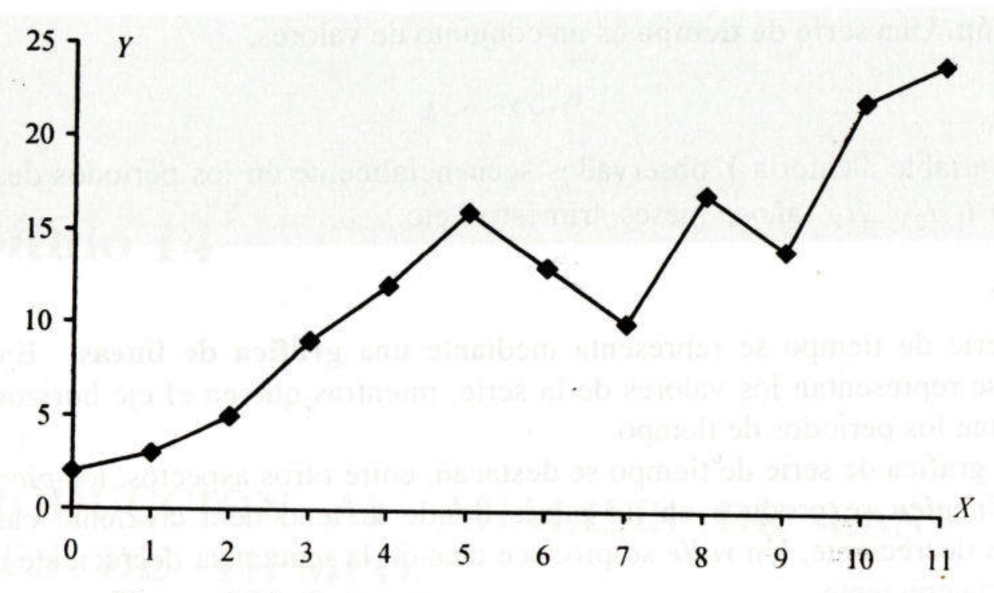


Figura. 14.1. *Gráfica de la serie de tiempo* del cuadro 14.1

14.2. Componentes de las series de tiempo.

Diversos tipos de cambios o movimientos están contenidos en una serie de tiempo. Estos movimientos son causados por factores o componentes, algunos de los cuales producen cambios a la serie a largo plazo, mientras que otros producen cambios a corto plazo (ver figura 14.2)

El análisis de la serie de tiempo es un proceso mediante el cual se llega a identificar y separar los factores o componentes que se relacionan con el tiempo y que influyen sobre los valores observados de la serie de tiempo.

Los factores que afectan a la serie de tiempo son cuatro, a saber:

1. **La *tendencia*** (o *tendencia secular*): Es el movimiento general creciente o decreciente de los valores de la serie de tiempo Y, que persiste en un periodo largo de tiempo. El más importante y básico es una línea recta. La componente de tendencia se denotará por T (Figura 14.2.a).

2. Las ***fluctuaciones cíclicas***: Son movimientos hacia arriba y hacia abajo de la línea de tendencia, y que ocurren en periodos cortos de tiempo. Se le llama así por que son secuencias repetidas del mismo modo que gira una rueda. La componente de fluctuación cíclica se denotará por C (Figura 14.2.b).

3. Las ***variaciones estacionales***: Se llama así a las oscilaciones en la extensión de un año y tiene mas o menos la misma forma año tras año. La periodicidad de las oscilaciones pueden ser incluso horarios, diarios, semanales, mensuales o trimestrales dependiendo de la naturaleza de la serie, pero no

duran mas de un año. Denotaremos a la componente de variación estacional por E (Figura 14.2.c)

4. Los ***movimientos irregulares***: Son movimientos con respecto a la tendencia que se deben a causas aleatorias o esporádicas (como huelgas, inundaciones, etc.) y por tanto no pueden adjudicarse a efectos estacionales o cíclicos. Denotaremos a la componente de variaciones irregulares por I



Figura. 14.2 *Movimiento secular, cíclico y estacional.*

Afortunadamente, las variaciones irregulares, carecen relativamente de importancia, por esta razón son considerados solamente como parte de las variaciones estacionales o de las fluctuaciones cíclicas.

Sin embargo al analizar las variaciones cíclicas o estacionales las causas esporádicas o específicas como: huelgas, terremotos, inundaciones, etc., que contribuyen a las variaciones periódicas, se les debe analizar como variaciones irregulares, siempre que sea posible.

14.3. Modelos con series de tiempo

Para estudiar una componente o para aislar una o más componentes de una serie original, se debe descomponer la serie. Para descomponer una serie de tiempo se supone que existe algún tipo de relación entre las cuatro componentes que contiene. En general, se supone que una serie de tiempo contiene sus componentes en forma aditiva y en forma multiplicativa.

El *modelo aditivo* supone que el valor de los datos originales, Y, es la suma de las cuatro componentes. Esto es,

$$Y = T + C + E + I$$

El *modelo multiplicativo* supone que el valor de los datos originales Y, es el producto de las cuatro componentes.

$$Y = T \times C \times E \times I$$

En el modelo aditivo, los datos se expresan en las unidades originales y el valor de una componente no afecta los valores de otros componentes. En el modelo multiplicativo, sólo la componente de tendencia se expresa en unidades originales, las componentes estacionales y cíclicas se expresan en números relativos o porcentajes, además, hay una dependencia mutua.

Por **ejemplo**, una producción de $Y = 37{,}800$ unidades (pares) de zapatos de una determinada empresa de calzado en el año 1999, se puede descomponer en $T = 40{,}000$ unidades, $C = 100\%$, que significa que no existe efecto debido al ciclo de negocios, $E = 105\%$, que significa que la producción de calzado tiene una variación estacional del $+5\%$ para ese año, $I = 90\%$, que significa que por algunas fuerzas no conocidas el número de zapatos producidos sufre una variación irregular de -10% en ese año. Entonces,

$$37{,}800 = 40{,}000 \times 1.00 \times 1.05 \times 0.90$$

El modelo multiplicativo es el que se utiliza más a menudo debido a que caracteriza a la mayoría de las series de tiempo económicas o de negocios. Trataremos, entonces con este modelo para separar las componentes que influyen en los valores de la serie de tiempo.

14.4. Análisis de la tendencia

El análisis de la tendencia es el procedimiento mediante el cual se determina la dirección del movimiento de la serie de tiempo a largo plazo y permite deducir el desarrollo de la serie en el futuro. Se supone que existe una tendencia y que esta puede ser: *ascendente*, *descendente*, *constante*. Lo primero que se debe decidir es si la tendencia es una línea recta o una curva.

La estimación de la tendencia se puede realizar por muchos métodos entre los que están: El *método de mano libre o alzada*, el *método de los dos promedios*, (o semipromedios), el método de las *medias móviles* y el método de los *mínimos cuadrados*.

Nos referiremos brevemente al método de mano libre y al de los dos promedios, para determinar la tendencia. El método de los promedios móviles será explicado en el proceso de determinación de los índices estacionales.

1) El método de mano libre,

Consiste en representar la serie de tiempo en un diagrama y después ajustar una línea recta a través de dos puntos del diagrama de modo que la recta represente la tendencia de la serie de tiempo.

2) El método de los dos promedios,

Consiste en dividir a la serie de tiempo en dos partes, se calcula la media de cada parte y se ajusta una línea de tendencia que pase por las dos medias.

EJEMPLO 14.2.

Continuando con el ejemplo 14.1. Determine la tendencia por el método de los dos promedios.

SOLUCION

Sea X, los años codificados en la forma: 1993, $X = 0$, 1994, $X = 1$,etc. La serie de datos se divide en dos partes iguales. La media de la primera parte que corresponde a 1993, 1994 y 1995 es $(5 + 7 + 9)/3 = 7$, esta media representa al año central 1994 ($X = 1$).

Además, la media de la segunda parte que corresponde a 1996, 1997 y 1998 es $(12 + 14 + 19)/3 = 15$, esta media representa al año central 1997 ($X = 4$)

Cuadro 14.2

Año	X	Y
1993	0	5
1994	1	7
1995	2	9
1996	3	12
1997	4	14
1998	5	19

En consecuencia, la recta: $Y = a + bX$, que buscamos, debe pasar por los puntos (1, 7) y (4, 15)

Entonces,

$$7 = a + b\ ,$$

$$15 = a + 4b$$

Resolviendo el sistema para a y b se obtiene $a = 3$ y $b = 4$.

La ecuación de tendencia es:

$$Y_T = 3 + 4X\ , \quad \text{(Origen: 1/7/1993, } X\text{: periodo 1 año)}$$

El método de mínimos cuadrados es el más usado por que es el mejor. Es el mejor por que se busca obtener una línea tal que la suma de los cuadrados de los desvíos o residuos e_i de la línea a los datos sea mínima.

Una serie de tiempo trata una cantidad variable Y en función del tiempo X. La unidad de tiempo por lo general es: 1 año, 1 mes o un trimestre.

La ecuación de tendencia no es una ecuación de regresión, por que la variable dependiente Y no es una variable aleatoria, y por que los valores de X en períodos de tiempo adyacentes son independientes

3) Tendencia de mínimos cuadrados

De todos los modelos posibles de tendencias de las series de tiempo el más importante es la línea recta.

Si Y_T representa los valores de la tendencia, la tendencia lineal es la expresión:

$$Y_T = a + bX$$

Las fórmulas para determinar los valores de a y b por el método de mínimos cuadrados son, como ya se vio en regresión lineal simple:

$$b = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$a = \overline{Y} - b\overline{X}$$

EJEMPLO 14.3.

Continuando con el ejemplo 14.1,

a) Determine la ecuación de tendencia lineal mediante el método de mínimos cuadrados.

b) Pronostique los valores de la serie para los años 1987–1998

c) Dibuje la línea de tendencia sobre la gráfica de la serie de tiempo.

SOLUCION.

Sea X los años codificados, donde $X = 0$, representa al año 1987, $X = 1$, representa al año 1988, además, representemos por Y las producciones en millones de unidades.

Observe que el origen de la tendencia puede elegirse en cualquier año (o periodo) de la serie de tiempo.

Los datos del ejemplo 14.1 re repiten en el cuadro 14.3. Donde, además se incluyen los cálculos para determinar la ecuación de la tendencia.

Cuadro 14.3. *Cálculo de la tendencia lineal*

Año	X	X^2	Y	XY	Y_T
1987	0	0	2	0	2.5
1988	1	1	3	3	4.3
1989	2	4	5	10	6.0
1990	3	9	9	27	7.8
1991	4	16	12	48	9.6
1992	5	25	16	80	11.4
1993	6	36	13	78	13.1
1994	7	49	10	70	14.9
1995	8	64	17	136	16.7
1996	9	81	14	126	18.5
1997	10	100	22	220	20.2
1998	11	121	24	264	22.0
Total	66	506	147	1062	

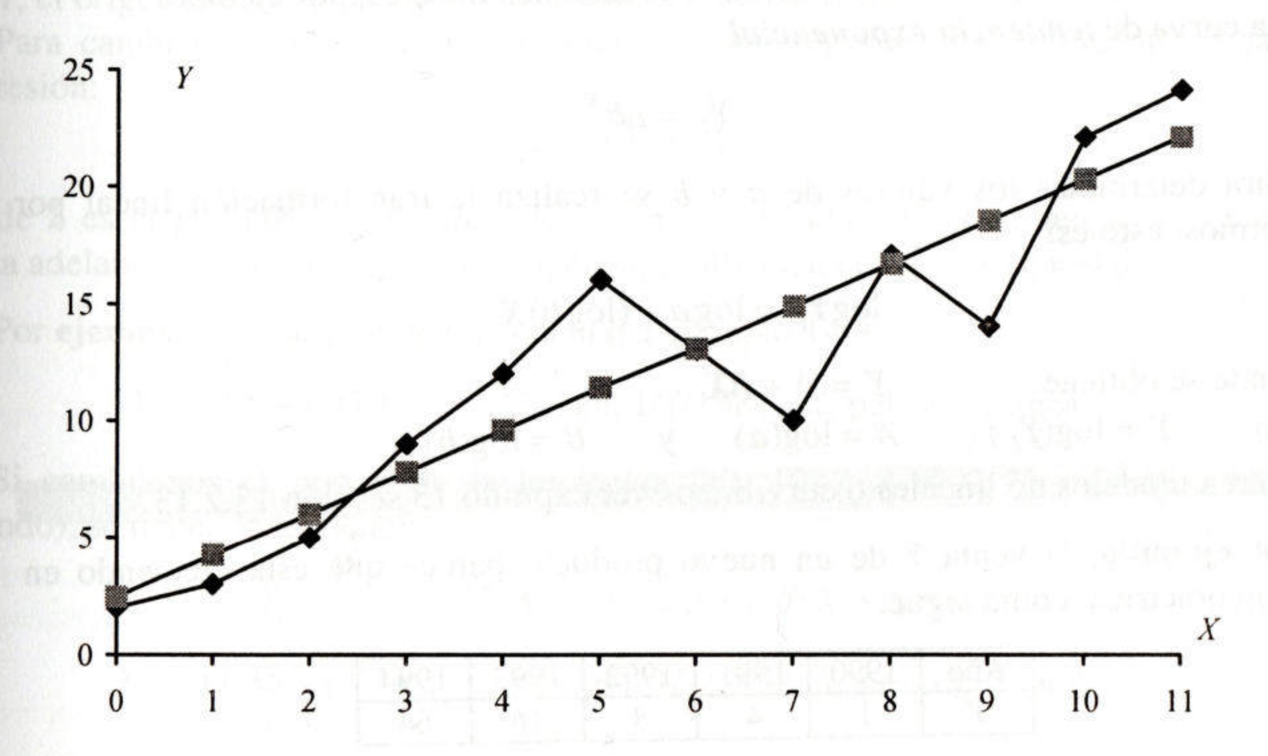


Figura 14.3: *Gráfica de la tendencia y la serie de tiempo*

Del cuadro 14.3 o utilizando el paquete *MCEST*, se obtienen:

$$\overline{X} = \frac{66}{12} = 5.5 \;,\; \overline{Y} = \frac{147}{12} = 12.25$$

$$b = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2} = \frac{1062 - 12(5.5)(12.25)}{506 - 12(5.5)^2} = \frac{253.5}{143} = 1.77$$

$$a = \overline{Y} - b\overline{X} = 12.25 - 1.77(5.5) = 2.5$$

$Y_T = 2.5 + 1.77X$, (Origen: 1/7/1987, X: periodo 1 año)

La pendiente: $b = 1.77$, indica que en los 12 años la producción ha tenido un aumento promedio de 1.77 millones de objetos anuales.

La gráfica de la tendencia y de la serie de tiempo es la figura 14.3

4) Curvas de tendencia

Si las tendencias no son lineales, con frecuencia se utilizan modelos no lineales o curvilíneos que se pueden transformar a ecuaciones lineales, por ejemplo:
La curva de ***tendencia exponencial***

$$Y_T = ab^X$$

Para determinar los valores de a y b se realiza la transformación lineal por logaritmos, esto es:

$$\log Y_T = \log a + (\log b)X$$

de donde se obtiene: $Y = A + BX$
siendo: $Y = \log(Y_T)$, $A = \log(a)$ y $B = \log(b)$.
Para otros modelos no lineales o curvilíneos ver capitulo 13 sección 13.2.13.

Por **ejemplo**, la venta Y de un nuevo producto parece que está creciendo en razón geométrica como sigue:

Año	1990	1991	1992	1993	1994
Y	1	4	8	16	64

Determine la ecuación de tendencia: $Y = ab^X$ y pronostique la producción para el año 1996.

Aplicando un paquete de computo, por ejemplo el paquete estadístico *MCEST* se obtiene:

La ecuación lineal transformada por log:

$$Y' = 0.006 + 0.42X$$

La ecuación exponencial:

$$Y_T = 1.1487(2.639)^X$$

Para el año 1996, X=6, se pronostica la venta $\hat{Y}$ = 388 unidades vendidas.

14.4.1 Cambio de origen de la tendencia y de la unidad de tiempo

1) Cambio de origen de la tendencia

Cuando se elige el origen de una serie de tiempo Y, el origen preciso es el centro del periodo inicial. Por ejemplo, la serie del ejemplo 14.1 se origina en el año 1987, el origen preciso es el centro de ese año, es decir 1° de Julio de 1987.

Para cambiar el origen de la tendencia de una serie de tiempo, se utiliza la expresión:

$$Y_T = a + b(X \pm k)$$

donde k es el número de unidades de tiempo cambiadas. Si el origen se cambia hacia adelante, k es positivo, y si el origen se cambia hacia atrás, k es negativo.

Por **ejemplo**, la ecuación de tendencia del ejemplo 14.3 es:

$$Y_T = 2.5 + 1.77X\ , \quad \text{(Origen, 1/7/1987, } X\text{, periodo 1 año)}$$

Si cambiamos el origen de la tendencia de 1987 a 1992 (5 unidades del periodo), se tiene, la tendencia:

$$Y_T = 2.5 + 1.77(X + 5) = 2.5 + 1.77X + 8.85$$

$$Y_T = 11.35 + 1.77X\ , \quad \text{(Origen, 1/7/1992, } X\text{, periodo 1 año)}$$

2) Reducción de valores de la tendencia anuales a mensuales y a trimestrales

La reducción de los valores de la tendencia anuales a valores mensuales se ejecuta utilizando la siguiente fórmula de transformación:

$$Y_T = \frac{a}{12} + \frac{b}{12} \times \frac{X}{12} = \frac{a}{12} + \frac{b}{144} X$$

La reducción de los valores de la tendencia anuales a valores trimestrales se ejecuta mediante la expresión:

$$Y_T = \frac{a}{4} + \frac{b}{4} \times \frac{X}{4} = \frac{a}{4} + \frac{b}{16} X$$

Por **ejemplo**, la ecuación tendencia anual:

$$Y_T = 2.5 + 1.77X \text{ (Origen: 1/7/1987, } X\text{, periodo 1 año)}$$

con origen, 1° de julio de 1987, unidad de tiempo 1 año y producción en millones de objetos, convertida en tendencia mensual es:

$$Y_T = \frac{2.5}{12} + \frac{1.77}{144} X = 0.208 + 0.0123X$$

$$Y_T = 0.208 + 0.0123X \text{ (Origen: 1/7/1987, } X\text{, periodo 1 mes)}$$

con origen 1° de julio de 1987, unidad de tiempo 1 mes y producción en millones de objetos.

Para tener el origen en el centro del periodo, es decir el 15 de julio de 1987, escribimos la ecuación:

$$Y_T = 0.208 + 0.0123(X + 0.5) = 0.21415 + 0.0123X$$

$$Y_T = 0.21415 + 0.0123X \text{ (Origen: 15/7/1987, } X\text{, periodo 1 mes)}$$

14.5. Análisis de las variaciones cíclicas

Los datos anuales contienen sólo dos componentes: la tendencia y el ciclo. Las variaciones estacionales son cambios mensuales o trimestrales que no se revelan en los datos anuales. Las variaciones irregulares, tienen efectos positivos y negativos en periodos cortos y tienden a compensarse en el curso de un año.

Por esta razón, cuando los datos son anuales, se pueden aislar los ciclos, suponiendo que el modelo multiplicativo es: $Y = T \times C$, y dividiendo luego los valores de la serie entre los correspondientes valores de la tendencia expresado en porcentajes. Esto es, los índices cíclicos están dados por:

$$\frac{Y}{Y_T} = \frac{T \times C}{T} = C$$

Una medida relativa de 100% indica ausencia de efectos cíclicos sobre el valor anual de la serie de tiempo.

EJEMPLO 14.4

a) Determine el componente cíclico de cada uno de los valores de la serie de tiempo del ejemplo 14.1, utilizando la ecuación de tendencia del ejemplo 14.3.

b) Represente gráficamente, los porcentajes de desviaciones de los relativos cíclicos obtenidos en a)

SOLUCION.

a) El cálculo de los índices cíclicos (o relativos cíclicos) se dan en el cuadro 14.4. La tercera columna contiene las estimaciones de la tendencia desde 1987 hasta 1998. En la cuarta columna, se dan los cocientes de valor observado entre el correspondiente valor estimado, expresado en porcentajes. La última columna contiene los desvíos con respecto al 100%.

b) En la figura 14.34 es la gráfica los índices cíclicos en porcentajes.

Cuadro 14.4. *Cálculo de índices relativos cíclicos.*

Año	Producción: Y	Tendencia: Y_T	Indices cíclicos: $(Y/Y_T)100$	% de desviaciones
1987	2	2.5	80.0	−20.0
1988	3	4.3	69.8	−30.2
1989	5	6.0	83.3	−16.7
1990	9	7.8	115.4	+15.4
1991	12	9.6	125.0	+25
1992	16	11.4	140.4	+40.4
1993	13	13.1	99.2	−0.8
1994	10	14.9	67.1	−32.9
1995	17	16.7	101.8	+1.8
1996	14	18.5	75.7	−24.3
1997	22	20.2	108.9	+8.9
1998	24	22.0	109.1	+9.1

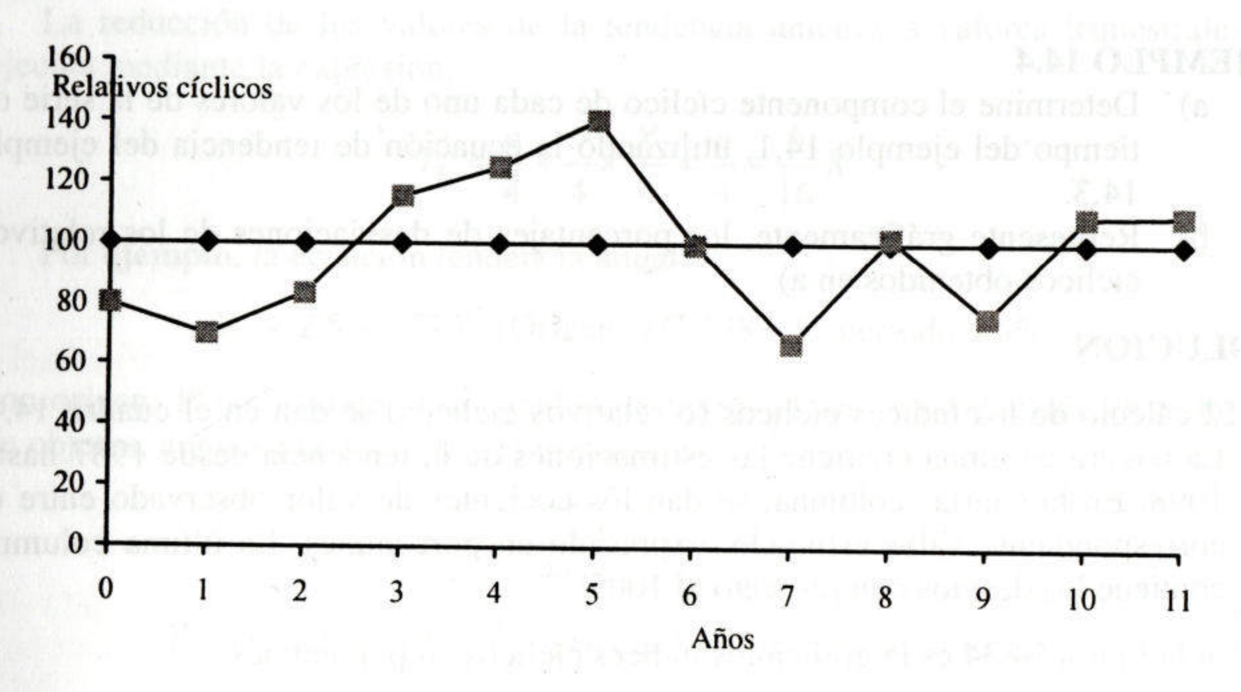


Figura 14.4. *Relativos cíclicos en porcentajes.*

14.6. Medición de variaciones estacionales

Las variaciones estacionales se calculan a partir de datos trimestrales (o mensuales). El efecto de la componente estacional de la serie de tiempo se identifica determinando el *número índice estacional* correspondiente a cada trimestre (o mes) del año. Por ejemplo, un índice estacional de 105% para un mes determinado, indica que los valores de la serie de tiempo para ese mes han promediado 5% por encima de los otros meses, debido a algún factor estacional positivo.

Existen varios métodos para determinar las índices estacionales de una serie de tiempo. El método que se utiliza con mayor frecuencia es el ***método del cociente (o la razón) de los promedios móviles***.

Para utilizar el método del cociente de promedios móviles, se supone que el modelo matemático de la serie de tiempo, es el modelo multiplicativo:

$$Y = T \times C \times E \times I$$

En primer lugar se determina la tendencia $T \times C$ empleando un **promedio móvil** de 12 meses (o de 4 trimestres.). Luego se elimina $T \times C$ para calcular $E \times I$ dividiendo los datos originales (Y) de la serie entre $T \times C$, esto es,

$$\frac{\text{Datos originales } Y}{\text{Media movil}} = \frac{T \times C \times E \times I}{T \times C} = E \times I$$

En segundo lugar se eliminan la componente irregular a partir $E \times I$ por un proceso de promediación, tal como se muestra con el siguiente ejemplo.

EJEMPLO 14.5

En el cuadro 14.5 que sigue, se presentan las producciones trimestrales (en cientos de miles de unidades) de la empresa ficticia cuya producción anual en millones de unidades se dan en la tabla del ejemplo 14.1.

Determine los índices estacionales de cada trimestre, utilizando el método del cociente del promedio móvil de 4 términos.

Cuadro 14.5. *Producción trimestral en cientos de miles de unidades.*

Año	1er. Trim.	2do. Trim.	3er. Trim.	4to. Trim.
1995	40	36	44	50
1996	32	30	36	42
1997	51	50	58	61
1998	57	55	63	65

SOLUCION.

El procedimiento de cálculo se presenta en la hoja de trabajo del cuadro 14.6 que sigue.

Cuadro 14.6. *Hoja de trabajo: producción trimestral en cientos miles de unidades*

(1) Años	(2) Trim.	(3) Datos originales $T \times C \times E \times I$	(4) Total móvil de 4 trimestres	(5) Total móvil centrado de 4 trimestres	(6) Prom. móvil Centrado $T \times C$ (5)/8	(7) Cociente del prom. móvil centrado (%) $E \times I$ (3)/(6)
1995						
	1	40				
	2	36				
			170			
	3	44		332	41.50	106.0
			162			
	4	50		318	39.75	125.8
			156			
1996						
	1	32		304	38.00	84.2
			148			
	2	30		288	36.00	83.3
			140			
	3	36		299	37.375	96.3
			159			
	4	42		338	42.25	99.4
			179			

1997	1	51		380	47.50	107.4
			201			
	2	50		421	52.625	95.0
			220			
	3	58		446	55.75	104.0
			226			
	4	61		457	57.125	106.8
			231			
1998	1	57		467	58.375	97.6
			236			
	2	55		476	59.50	92.4
			240			
	3	63				
	4	65				

En la tercera columna (3) los datos originales Y, contienen los factores: T, C, E, I, en el modelo supuesto, $Y = T \times C \times E \times I$

En la cuarta columna (4), el primer total móvil de 4 trimestres, 170, (suma de las ventas de los 4 primeros trimestres del año 1995), se *coloca sobre la línea que separa* el segundo y tercer trimestre del año 1995. El segundo total móvil, 162, (total de los trimestres 2, 3, y 4 de 1995 y del trimestre 1 del año 1996) se coloca sobre la línea que separa el trimestre 3 y 4 de 1995, y así sucesivamente sobre los otros totales.

En la quinta columna (5), se colocan los totales móviles *centrados* en cada trimestre. El total móvil centrado en cada trimestre se obtiene sumando los dos totales móviles que se encuentran juntos y que involucra a 8 trimestres o dos años.

En la sexta columna (6) se colocan las *promedios móviles centrados,* $T \times C$, de cuatro trimestres. Cada media móvil es igual al total móvil centrado, dividido entre 8. **Estos promedios son las tendencias por el método de los promedios móviles.**

En la columna (7) se colocan el *cociente del promedio móvil,* $E \times I$, en porcentaje. Cada cociente de promedio móvil se obtiene dividiendo el dato original del trimestre entre el promedio móvil centrado en ese trimestre y expresado en porcentaje.

Una vez obtenido, $E \times I$ (promedio libre de tendencia y de componente cíclica), se puede obtener E calculando el promedio de los, $E \times I$

Para obtener finalmente los índices estacionales, los valores de $E \times I$ se obtienen de la última columna (7) de los datos de la hoja de trabajo anterior.

En cada trimestre se promedian estos porcentajes dando como resultado *los índices estacionales no ajustados* trimestrales.

Los *índices estacionales* trimestrales se obtienen multiplicando cada uno de los promedios o índices estacionales no ajustados por **el factor trimestral de ajuste** que está dado por:

$$\text{Factor trimestral de ajuste} = \frac{400}{\text{suma de medias trimestrales}}$$

Hoja de trabajo: *Cálculo de Indices estacionales trimestrales*

Año	1er. Trim.	2do. Trim.	3er. Trim.	4to. Trim.	Total
1995			106.0	125.8	
1996	84.2	83.3	96.3	99.4	
1997	107.4	95.0	104.0	106.8	
1998	97.6	92.4			
Indice sin Ajustar	96.4	90.3	102.1	110.7	399.4
Indices Estacionales	96.5	90.4	102.3	110.9	

Factor de ajuste = 400/399.4=1.0015

EJEMPLO 14.6.

Con referencia a la tabla 14.5 del ejemplo 14.5, realice el ajuste estacional de los datos.

SOLUCION.

Los valores ajustados estacionalmente se muestran en la tabla que sigue. Cada uno de estos valores (desestacionalizados) se ha obtenido dividiendo el valor trimestral de cada año entre el correspondiente índice estacional para este trimestre y multiplicando por 100.

Por **ejemplo**, el valor 41.45 del primer trimestre de 1995 se ha obtenido dividiendo el valor 40 (valor del primer trimestre de 1995) entre el índice estacional correspondiente 96.5 y multiplicando por 100, también, 33.16=(32/96.5)×100 Así mismo, el valor 39.82 del segundo trimestre de 1995, se ha obtenido dividiendo 36 entre 90.4 (índice estacional del segundo trimestre de 1995) y multiplicado por 100, etc.

***Producción trimestral** (en cientos de miles de unidades)*
ajustados estacionalmente

Año	1er. Trim.	2do. Trim.	3er. Trim.	4to. Trim.
1995	41.45	39.82	43.01	45.09
1996	33.16	33.19	35.19	37.87
1997	52.85	55.34	56.70	55.00
1998	59.06	60.84	61.58	58.61

EJEMPLO 14.7
Con referencia a la ecuación de tendencia anual del ejemplo 14.3

a) Convierta a una ecuación de tendencia en términos trimestrales con origen en la mitad del primer trimestre de 1987 y expresarlo en cientos de miles de unidades.

b) Pronostique el nivel de la producción trimestral para cada uno de los cuatro trimestres del año 1999 con base a los índices estacionales obtenidos en el ejemplo 14.5.

SOLUCION.

a) La ecuación de tendencia de la producción anual en millones de unidades es:

$$Y_T = 2.5 + 1.77X\,, \quad \text{(Origen, 1/7/1987, } X\text{, periodo 1 año)}$$

La conversión trimestral es:

$$Y_T = \frac{a}{4} + \frac{b}{16}X = \frac{2.5}{4} + \frac{1.77}{16}X = 0.625 + 0.1106X$$

$$Y_T = 0.625 + 0.1106X \quad \text{(Origen, 1/7/1987, } X\text{, periodo trimestral)}$$

Cambiando el origen a la mitad del primer periodo de 1987, ($k = -1.5$), se tiene:

$$Y_T = 0.625 + 0.1106(X - 1.5) = 0.4591 + 0.1106X$$

$$Y_T = 0.4591 + 0.1106X \text{ en millones de unidades}$$

(Origen, 15/02/1987, X, periodo trimestral)

La tendencia expresada en cientos de miles de unidades es:

$$Y_T = 10(0.4591 + 0.1106)X = 4.591 + 1.106X$$

(Origen, 15/02/1987, X, periodo trimestral)

b) Dado que el año 1987, origen de la serie trimestral, corresponde a X=0, entonces, los trimestres 1, 2, 3 y 4 del año 1999, corresponden a los valores X=48, X=49, X=50, X=51, respectivamente. Entonces, los valores estimados en cada trimestre del año 1999, aplicando la ecuación trimestral de tendencia y ajustados por los índices estacionales trimestrales son:

Primer trimestre de 1999 = [4.591+1.106(48)](96.5/100) = 55.66
Segundo trimestre de 1999 = [4.591+1.106(49)](90.4/100) = 53.14
Tercer trimestre de 1999 = [4.591+1.106(50)](102.3/100) = 61.27
Cuarto trimestre de 1999 = [4.591+1.106(51)](110.96/100) = 67.68

EJERCICIOS

1. Dados los siguientes datos demográficos,

Año	Población (en millones)	año	Población (en millones)
1985	50	1991	64
1986	52	1992	67
1987	55	1993	70
1988	58	1994	74
1989	60	1995	78
1990	62	1996	80

a) Grafique la serie de tiempo.

b) Determine y trace la línea de tendencia por el método de mano alzada.

c) Determine la línea de tendencia por el método de los semipromedios.

Rp. c) puntos (2.5,56) y (8.5,72), 1985, X=0, $\hat{Y} = 49+2.67X$, Origen: 1/7/1985, X: periodo 1 año

2. Dados los siguientes datos sobre la producción de una compañía en miles de unidades

Año	1992	1993	1994	1995	1996	1997
Producción	15	14	18	20	17	24

a) Grafique la serie de tiempo.

b) Halle la ecuación de tendencia lineal con origen: 1/7/1992, X: periodo 1 año.

c) Estime la producción para cada año. Calcule la componente cíclica de la serie y grafique en un solo sistema: la tendencia, la componente cíclica.

d) Cambie la ecuación de tendencia de ventas anuales a una ecuación de tendencia de ventas mensuales con base al 15 de enero de 1992, y estime la producción para marzo de ese año.

Rp. b) 1992, X=0 $\hat{Y} = 14 + 1.6X$, Origen: 1/7/1992, X: periodo 1 año, c) 107.14, 89.74, 104.65, 106.38, 83.33, 109.09 d) $\hat{Y} = 1.167 + 0.011X$, origen: 1/7/1992, periodo 1 mes,
$\hat{Y} = 1.1065 + 0.011X$, origen: 15/1/1992, periodo 1 mes.

3. El total de ventas anuales, en miles de dólares, de la empresa G&P se dan en la siguiente tabla:

Año	1992	1993	1994	1995	1996	1997	1998
Ventas	8	12	17	18	20	23	25

a) Ajuste una recta de tendencia por el método de mínimos cuadrados.

b) Calcule los valores de la tendencia para cada año

c) Suponga que la tendencia continúe para cada año, ¿cuál será la tendencia para 2001.

d) ¿En que año diría usted que la venta alcanzará los $39,286?

Rp. a) $\hat{Y} = 9.43+2.7X$, c) $33.73, d) en el año 2003..

4. Dada la siguiente ecuación de tendencia de 1983 a 1996

$$Y_T = 94.2 + 1.28X$$

Origen, 1/7/ 1993, *X:* periodo 1 año

donde *Y* es la producción de un bien en millones de kilogramos.

a) Cambie el origen a 1/7/1983, *X*, periodo 1 año.]
b) Cambie el origen a 1/7/1996, *X*: periodo 1 año.
c) Cambie el origen a 1/1/1990, *X*: periodo 1 año.
d) Cambie el origen a 1/10/1997, *X*: periodo 1 año.

Rp. Y_T=94.2 + 1.28($X + k$), a) k=−7, b) k=6, c) k=1/2, d) k=1+1/4=5/4

5. Dada la siguiente ecuación de tendencia de 1983 a 1996

$$Y_T = 3{,}600 + 120X \qquad \text{Origen, 1/7/ 1990, } X\text{: periodo 1 año}$$

donde *Y* es la producción de un bien en cientos de toneladas.

a) Convierta la tendencia anual a tendencia mensual con origen 1/7/1990
b) Convierta la tendencia anual a tendencia mensual con origen 15/1/1990
c) Convierta la tendencia anual a tendencia trimestral con origen 1/7/1990
d) Convierta la tendencia anual a tendencia trimestral con origen 1/1/1990
e) Cambie el origen a 1/1/1990, *X*: periodo 1 año.
f) Cambie el origen a 1/10/1997, *X*: periodo 1 año.

Rp. Y_T=94.2 + 1.28($X + k$), a) k=−7, b) k=6, c) k=1/2, d) k=1+1/4=5/4

6. Dada la siguiente ecuación de tendencia

$$Y_T = 1{,}500 + 36X$$

Origen, 1/7/ 1990, *X*: periodo 1 año

a) Interprete la ecuación de tendencia.
b) Convierta la ecuación a períodos mensuales y con origen en Junio 15 de 1990.

7. Se ajustó una recta de tendencia para el ingreso anual de los empleados de una empresa, durante 5 años, 1992-1996, por el método de los mínimos cuadrados.
a) La ecuación mostró que en 1992 el ingreso sería \$1,040, y de \$3,920 en 1996. Encuentre la ecuación con el origen en 1993
b) ¿Cuál es el ingreso para 1994?
c) Cambie esta ecuación anual por una ecuación mensual con origen en 15 de Julio de 1993.
d) ¿Cuál es la cantidad total de ingreso que el empleado de esta empresa ha ganado durante los cinco años comprendidos entre 1992-1996?

8. Determine la recta de tendencia del problema 1 por el método de los mínimos cuadrados y calcule el componente cíclico para cada uno de los valores de la serie de tiempo.

Rp. a) $\hat{Y} = 49.17+2.73X$, b) 101.7, 100.2, 100.7, 101.1, 99.9, 98.7, 97.7, 98.2, 98.6, 100.4, 102.0, 101.1.

9. Utilizando la ecuación de tendencia del problema 2, determine el componente cíclico de cada uno de los valores de la serie de tiempo.

Rp. 107.1, 89.7, 104.7, 106.4, 83.3, 109.1.

10. En la tabla que sigue se presentan los datos trimestrales del número de pedidos (en millares) de cierto tipo de objeto que se recibieron en una empresa ficticia.

Número de pedidos en una empresa

Año	1er. Trim.	2do. Trim.	3er. Trim.	4to. Trim.
1993	10	14	17	22
1994	11	15	16	19
1995	12	20	18	22
1996	11	13	17	19
1997	15	17	25	27
1998	19	21	28	29

Encuentre los pedidos totales por año y encuentre la ecuación de tendencia lineal por el método de mínimos cuadrados y en base en el año 1993.

Rp. $\hat{Y} = 56.62+6.5X$,

11. Continuando con el problema 10 y la ecuación de tendencia obtenida, determine el componente cíclico de cada uno de los valores de la serie.

Rp. 111.3, 96.7, 103.5, 78.9, 101.7, 108.9.

12. Continuando con el problema 10 determine los índices estacionales trimestrales.

Rp. 74.7, 93.5, 107.7, 124.2.

13. Para los datos que siguen sobre producción (en millones de unidades),

Año\mes	E	F	M	A	M	J	J	A	S	O	N	D
1994	5	5	7	7	9	10	8	8	7	5	5	5
1995	6	5	7	6	9	11	7	8	7	5	4	4
1996	7	4	8	7	10	12	6	7	8	6	7	5
1997	8	5	7	8	10	13	8	7	8	6	7	6
1998	4	6	9	8	11	14	9	8	9	7	9	5

a) Determine la tendencia anual por el método de los mínimos cuadrados. (convierta primero los datos mensuales a anuales).

b) Calcule los índices cíclicos anuales en cada uno de los valores de la serie

Rp. a) $\hat{Y} = 77.8 + 5X$, b) 104.1, 95.4, 99.1, 100.2, 101.2

14. Con los datos del problema 13 determine los índices estacionales trimestrales.

Rp. 86.8, 133.8, 104.6, 74.8.

15. Con los datos del problema 13, calcule la tendencia por el método de promedio móvil de 4 trimestres.

Rp. 112.9, 73.2, 88.3, 130, 110.7, 63.8, 92.1, 137.3, 96, 80.9, 87.9, 134.1, 99.5, 81.7, 79.6, 134.7.

Capítulo 15

INTRODUCCION A LAS PRUEBAS ESTADISTICAS NO PARAMETRICAS.

Introducción

Las pruebas hipótesis en las que se han utilizado la estadística z, la t o la F se denominan *métodos paramétricos*. En estas pruebas, se supone que la distribución de probabilidad de la población de la cual se extrae la muestra tiene determinada forma y que su (o sus) parámetro verifica ciertas condiciones de manera que la estadística muestral correspondiente tenga una distribución de probabilidad conocida en la que se puede determinar una región crítica y establecer una regla de decisión. Por ejemplo, para la aplicación de la prueba t, se requiere suponer que la población es normal.

Además, para la aplicación de las pruebas paramétricas se requiere que el nivel de las mediciones muestrales sea al menos de escala de intervalos.

Sin embargo, existen muchas aplicaciones en las ciencias e ingeniería donde no es posible conocer la distribuciones de la poblaciones de las que se extraen las muestras o donde los datos se reportan como *valores en escala ordinal*. En estos casos, se utilizan métodos alternativos equivalentes a los paramétricos denominados **métodos no paramétricos** o **de distribución libre**.

Con frecuencia se utilizan las pruebas no paramétricas cuando se tratan de inferencias con muestras pequeñas y distribución desconocida de la población, ya que en estos casos no se puede utilizar el teorema central del límite. La aplicación de los métodos no paramétricos no requiere conocimientos matemáticos avanzados, debido a que la tarea matemática consiste en general, ordenar por rangos los datos observados.

Si se verifican las condiciones exigidas para el uso de una prueba paramétrica, tal como la prueba t por ejemplo, entonces, es siempre preferible utilizar ésta y no su equivalente no paramétrico. Esto se debe a que si se utiliza el mismo nivel de significación en ambas pruebas, la potencia de una prueba no paramétrica es

siempre menor a la de su equivalente paramétrica (ver por ejemplo [1] página 465). Por otro lado, con los métodos no paramétricos se pierde gran cantidad de información al no operar explícitamente con las puntuaciones si no con sus rangos o con el modo como estas se agrupan en categorías.

Si son aplicables tanto la prueba paramétrica como la no paramétrica al mismo conjunto de datos es pues preferible aplicar la prueba paramétrica.

Las prueba de hipótesis con la estadística chi-cuadrado, entre ellas **bondad de ajuste y tablas de contingencia, son pruebas no paramétricas**. Debido a la variedad de sus aplicaciones se dedicó el capitulo 11 de este texto a las pruebas con chi-cuadrado.

Los ejemplos y ejercicios han sido resueltos utilizando el paquete estadístico didáctico *MCEST* creado por el autor de este libro.

15.1 Algunas técnicas no paramétricas

15.1.1 Una sola muestra: La prueba de Kolmogorov-Smirnov

El **método de Kolmogorov-Smirnov** es un procedimiento no paramétrico que se utiliza para comprobar la hipótesis nula de que la muestra procede de una población en la que la variable está distribuida según la **normal** (también según la uniforme o Poisson).

La estadística de la prueba se denota por D y se define por:

$$D = \text{Máxima} \mid F_0(x_i) - S_n(x_i) \mid .$$

Donde $F_0(x_i)$ y $S_n(x_i)$ son las probabilidades acumuladas esperadas y observadas respectivamente, tal como **se explica en el ejemplo 11.5 del capitulo 11 de este texto**.

En la tabla de Kolmogorov-Smirnov, se encuentran algunos valores críticos de la distribución muestral de D para diversos valores de n y α dados. Se rechaza la hipótesis nula que los datos provienen de una distribución normal contra la alternativa de que los datos no provienen de la distribución normal, si el valor de D es mayor que el valor crítico correspondiente. No se rechaza la hipótesis nula en caso contrario.

En el **ejemplo** 11.5, **la diferencia** máxima en valor absoluto de la frecuencias observadas y esperadas es **0.0967**.

Por otra parte en la tabla de valores de Kolmogorov-Smirnov para el nivel de significación $\alpha = 0.01$, y $n = 10$ se encuentra el valor crítico 0.490. Dado que $0.0967 < 0.490$, deberíamos aceptar la hipótesis nula de que es normal la población de la cual se ha obtenido la muestra.

15.1.2 Una sola muestra: La prueba de los signos

Los métodos que se utilizan para probar la hipótesis nula que consiste en afirmar que la media poblacional μ es igual a un valor dado μ_0, esto es, $H_0: \mu = \mu_0$, contra una alternativa bilateral o unilateral son válidos sólo si la población es normal o si el tamaño de la muestra es $n \geq 30$.

Se requiere además que los valores de la muestra se encuentren cuando menos en escala de intervalos. Pero, si la población no es normal y si $n < 30$ o cuando los valores de la muestra se encuentran cuando menos en escala ordinal, se debe utilizar la prueba no paramétrica correspondiente.

En este texto se exponen dos métodos de prueba acerca del promedio utilizando una muestra: **La prueba de los signos y la prueba de Wilcoxon**.

En muchos procedimientos no paramétricos que se refieren a promedios, la media se reemplaza por la mediana como parámetro de prueba. Es evidente que si la distribución es simétrica la media y la mediana poblacionales son parámetros iguales.

La prueba de los signos se utiliza para probar la hipótesis nula de que el **promedio poblacional es igual a un valor dado**. Para aplicar la prueba del signo no se requiere hacer suposiciones de normalidad de la población, se requiere sí que los valores de la muestra se encuentren cuando menos en escala ordinal.

La hipótesis nula consiste pues en afirmar que la mediana de la población, parámetro que denotaremos por $\tilde{\mu}$ es igual a algún valor dado $\tilde{\mu}_0$, esto es: $H_0: \tilde{\mu} = \tilde{\mu}_0$. La alternativa puede ser de un extremo o de dos extremos.

El resumen del procedimiento de la prueba de los signos es como sigue:

Se asigna el signo "+" a cada valor mayor que $\tilde{\mu}_0$ se asigna el signo "–" a cada valor de la muestra menor que $\tilde{\mu}_0$. Se descarta la medición que es igual a $\tilde{\mu}_0$.

El número de valores no descartados es el tamaño *n* de la muestra.

Si la hipótesis nula es verdadera, entonces, debemos tener aproximadamente el mismo número de signos + y de signos −. Es decir; que la proporción de signos positivos (y de signos negativos) debe ser 0.5.

Si X representa **el número de veces que ocurre el signo menos frecuente**, entonces, X es una variable aleatoria discreta cuya distribución de probabilidades es binomial $B(n, p)$ con $p = 1/2$. Esta es la distribución que se utiliza en la prueba de los signos.

Si la prueba es unilateral de cola izquierda, esto es:

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \text{ contra } H_1: \tilde{\mu} < \tilde{\mu}_0$$

o si la prueba es unilateral cola a la derecha, esto es:

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \text{ contra } H_1: \tilde{\mu} > \tilde{\mu}_0$$

se rechazará H_0 si el valor P

$$P = P[X \leq x \text{ cuando } p = 1/2] \text{ donde } x < n/2$$

es menor que el nivel de significación α.

Por **ejemplo**, cuando $n = 10$, $x = 1$ en la tabla de probabilidades binomial se encuentra:

$$P = P[X \leq 1 \text{ cuando } p = 1/2] = 0.01074$$

de tal forma que se rechazará la hipótesis nula $H_0: \tilde{\mu} = \tilde{\mu}_0$ al nivel 5% pero no al nivel 1%.

Si la prueba es bilateral

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \text{ contra } H_1: \tilde{\mu} \neq \tilde{\mu}_0$$

se rechazará H_0 si el valor P calculado

$$P = 2P[X \leq x \text{ cuando } p = 1/2] \quad \text{si } x < n/2$$

es menor que el nivel de significación α.

Para una prueba bilateral con $\alpha = 0.05$, se requiere que el tamaño n de la muestra debe ser por lo menos 6 (ver referencia 17 página 488).

NOTA. Cuando $n \geq 30$ y $p = 1/2$, puede utilizarse la distribución normal como aproximación de la binomial (basta con $n > 10$ para tener una buena aproximación, ya que así se verifica $np > 5$).

La estadística de la prueba es en este caso:

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - n/2}{\frac{1}{2}\sqrt{n}}$$

cuya distribución es aproximadamente normal $N(0,1)$.

EJEMPLO 15.1.

Los siguientes datos representan el número de unidades vendidas en una muestra de 13 tiendas de un nuevo artículo de aseo personal:

15, 25, 32, 20, 34, 30, 31, 16, 28, 22, 36, 23

No se hacen suposiciones de normalidad de la distribución. Utilice la prueba de los signos al nivel de significación 5% para probar la hipótesis nula de que la mediana de las ventas no es mayor que 20 unidades.

SOLUCION.

Se asignan los signos "+" o "−" a cada valor de la muestra que es mayor o menor a 20 respectivamente y se asigna el 0 a la medición que es igual a 20, resultando la secuencia:

$$- + + 0 + + + - + + + +$$

para la cual $n = 11$ (signos, se descarta el 0), $x = 2$ (el número de veces que ocurre el signo menos frecuente)

Sea X el número de signos negativos de los $n = 11$ signos. La prueba continua:

1. *Hipótesis*: $H_0: \tilde{\mu} = 20$ contra $H_1: \tilde{\mu} > 20$

2. *Nivel de significación*: $\alpha = 0.05$

3. *Estadística*: Variable X binomial *con*, $n = 11$ y $p = 1/2$.

4. *Cálculos*: El valor calculado de P para esta prueba unilateral es:

$$P = P[X \leq 2 \text{ cuando } p = 1/2] = \sum_{k=0}^{2} C_k^{11}(0.5)^{10} = 0.0327.$$

6. *Decisión*. Dado que $P = 0.0327 < 0.05$, se rechaza H_0 y se acepta H_1.

NOTA. Al utilizar la aproximación de la curva normal, para $n = 11$, y $x = 2$ signos positivos se tiene:

$$np = 11(0.5) = 5.5$$

$$\sqrt{npq} = \sqrt{11(0.5)(0.5)} = 1.658$$

$$Z = \frac{2.5-5.5}{1.658} = -1.81$$

Por lo tanto,
$$P = P[X \leq 2] \cong P[Z < -1.81] = 0.0352.$$

lo cual conduce a rechazar la hipótesis nula en esta prueba unilateral al nivel de significación del 5%.

15.1.3. Una sola muestra: La prueba de rango con signo de Wilcoxon

Igual que en la prueba del signo, la prueba de rangos con signo de Wilcoxon conocida también como prueba T de Wilcoxon se utiliza para probar la hipótesis nula que el valor de una mediana poblacional es igual a un valor dado. Esto es, $H_0: \tilde{\mu} = \tilde{\mu}_0$ contra una alternativa unilateral o bilateral. Para aplicar la prueba, los valores de la muestra deben estar dados cuando menos en escala ordinal. No se requieren hacer suposiciones acerca de la forma de la distribución de la población.

En el caso de tratarse de una **distribución simétrica continua** se puede probar la hipótesis nula que la media poblacional es igual a un valor dado, esto es, $H_0: \mu = \mu_0$ contra cualquier alternativa unilateral o bilateral.

El procedimiento de la prueba es como sigue:

Se determina la diferencia entre cada uno de los valores observados y el valor hipotético de la mediana, digamos, $\tilde{\mu}_0$ Esto es, $d = X - \tilde{\mu}_0$

Si alguna de las diferencias es igual a cero, se elimina la observación correspondiente del análisis y de esta manera se reduce el tamaño efectivo de la muestra. El tamaño n de la muestra es el número de diferencias no nulas.

Después se ordenan los valores absolutos de la diferencia, de menor a mayor, asignando el rango 1 a la diferencia absoluta más pequeña, 2 a la siguiente diferencia menor, etc. Cuando las diferencias absolutas son iguales, se asigna el rango promedio a los valores que son iguales.

Finalmente, se obtienen por separado la suma de los rangos para las diferencias positivas y negativas.

Se rechazará, H_0: $\tilde{\mu} = \tilde{\mu}_0$ en favor de H_1: $\tilde{\mu} > \tilde{\mu}_0$, sólo si la suma de rangos "+", es mayor que la suma de rangos "–". De la misma forma se rechazará $H_0: \tilde{\mu} = \tilde{\mu}_0$ para aceptar H_1: $\tilde{\mu} < \tilde{\mu}_0$ sólo si la suma de rangos "–" es mayor que la suma de rangos "+". Finalmente, Se rechazará H_0: $\tilde{\mu} = \tilde{\mu}_0$ en favor de $H_1: \tilde{\mu} \neq \tilde{\mu}_0$ si ya sea la suma de rangos "+", o la suma de rangos "–" es pequeña.

Luego, cualquiera sea la hipótesis alternativa, se rechazará la hipótesis nula $H_0: \tilde{\mu} = \tilde{\mu}_0$ para la menor de las sumas de rangos "+" y "–". La menor de las dos sumas en valor absoluto designada por *T* es la estadística *T* de Wilcoxon.

Para el tamaño *n* de la muestra y niveles de significación α de 0.01, 0.025, 0.05 para prueba unilateral y 0.02, 0.05, 0.10 para pruebas de dos colas, en la tabla de valores de la *T* de Wilcoxon se encuentra el valor crítico *K*. Se rechaza la hipótesis nula H_0 si el valor de $T \leq K$. No se rechaza la hipótesis nula en caso contrario.

Para una prueba bilateral y valores pequeños del tamaño *n* de la muestra, se tienen los siguientes valores críticos para $\alpha = 0.05$ y $\alpha = 0.01$ respectivamente:

$$T_{\alpha=0.05} = \frac{n^2 - 7n + 10}{5}$$

$$T_{\alpha=0.01} = \frac{11n^2}{60} + 5 - 2n$$

Cuando $n \geq 30$ (bastaría con $n > 10$) y la hipótesis nula es cierta la estadística *T* tiene una distribución aproximadamente normal con:

$$\text{Media}: \quad \mu_T = \frac{n(n+1)}{4}, \text{ y}$$

$$\text{Varianza}: \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

En este caso, se puede utilizar la estadística:

$$Z = \frac{T - \mu_T}{\sigma_T}$$

para determinar la región crítica de la prueba.

EJEMPLO 15.2.

Continuando con el ejemplo 15.1, utilice la prueba de Wilcoxon para probar la hipótesis nula de que la mediana de las ventas es igual a 20 unidades, contra una alternativa bilateral. Utilice nivel de significación 5%.

SOLUCION.

Los datos del ejemplo 15.1 son:

15, 25, 32, 20, 34, 30, 31, 16, 28, 22, 36, 23

Al restar la mediana 20 de cada uno de los 12 valores de la muestra, se obtienen las diferencias una de las cuales es igual a cero. Descartando la diferencia nula, se obtiene: $n = 11$.

Los rangos que se asignan a las diferencias sin considerar el signo, se dan en la tabla que sigue (se aplicó el paquete *MCEST*)

Valores X	$d_i = x_i - \tilde{\mu}$	Rango +	Rango –
15	–5		– 4.5
25	+5	4.5	
32	+7	9.0	
20	0	–	–
34	+14	10.0	
30	+10	7.0	
31	+11	8.0	
16	–4		– 3.0
28	+8	6.0	
22	+2	1.0	
36	+16	11.0	
23	+3	2.0	
Total		58.5	7.5

Continuando con la prueba, se tiene:

1. *Hipótesis*: $H_0: \tilde{\mu} = 20$ contra, $H_1: \tilde{\mu} \neq 20$
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística* y región crítica de la prueba: T Wilcoxon. Para $n = 11$ y una prueba bilateral con $\alpha = 0.05$, en la tabla de valores críticos de Wilcoxon, se halla el valor 11. Se rechazará H_0 si $T \leq 11$
4. La suma de rangos negativos es 7.5 y la suma de rangos positivos es 58.5. La estadística de Wilcoxon para esta prueba bilateral es la menor de las dos sumas, luego, $T = 7.5$

5. *Decisión*. Dado que $T = 7.5 < 11$, se rechaza H_0 en esta prueba bilateral.

NOTA. Al utilizar la aproximación de la curva normal, para la muestra aleatoria de tamaño $n = 11$ y $T = 7.5$; se tiene:

$$\mu = \frac{n(n+1)}{4} = \frac{11(12)}{4} = 33$$

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{11(12)(23)}{24}} = 11.247$$

$$Z = \frac{7.5 - 33}{11.247} = -2.27$$

Por lo tanto:

$$P = P[Z \leq -2.27] = 0.0116.$$

Esta probabilidad P nos conduce a rechazar la hipótesis nula al nivel de significación del 5%.

15.1.4 Dos muestras dependientes: La prueba de los signos.

La técnica paramétrica que se utiliza para analizar los datos de dos muestras dependientes o correlacionadas es aplicar la prueba *t* a las diferencias de los valores observados en parejas. Para aplicar el método se supone que las diferencias están distribuidos en forma normal y que la medición de los datos de las muestras dependientes sea por lo menos en escala de intervalos.

Cuando no se cumplen los requisitos de normalidad o de mediciones por lo menos en escala de intervalos una prueba alternativa es la **prueba no para métrica de los signos**.

La prueba de los signos es un método no paramétrico que se aplica cuando se tienen **dos muestras relacionadas** para probar la hipótesis nula de que las distribuciones de las dos poblaciones (X,Y) son iguales (tienen medianas iguales o medias iguales si las poblaciones son continuas). No se requieren hacer suposiciones de normalidad de las dos distribuciones poblacionales. Sólo se requiere que los valores de las dos muestras aleatorias apareadas sean cuando menos de escala ordinal.

Específicamente, la hipótesis nula puede expresarse en dos formas:

1) $H_0: \mu_1 = \mu_2$, las dos medias poblacionales son iguales, siempre que se pueda suponer que las distribuciones de X y de Y son simétricas.

2) La mediana de las diferencias $x_i - y_i$ es cero, para n datos (x_i, y_i) observados en parejas en las poblaciones X e Y.

El resumen del procedimiento para la prueba de los signos es como sigue:

Se asigna el signo "+" a cada una de las parejas donde $x_i > y_i$, se asigna el signo "−" a cada una de las parejas donde $x_i < y_i$. Las parejas en que $x_i = y_i$ se eliminan del análisis. El número de parejas no descartadas es el **tamaño *n* de la muestra**.

Si la hipótesis nula es verdadera, es decir, si las dos poblaciones tienen igual mediana, debemos tener aproximadamente el mismo número de signos + y dė signos −. Es decir, que la proporción de signos positivos (y de signos negativos) debe ser igual a 0.5. Además, si X representa **el número de veces que ocurre el signo menos** frecuente, entonces, X una variable discreta cuya distribución es binomial $B(n, p)$ con p=1/2, la misma que se utiliza para realizar la prueba.

Para una prueba bilateral con $\alpha = 0.05$, se requiere que el tamaño n de la muestra sea por lo menos 6 (ver referencia 17 página 488).

Cuando $n \geq 30$ y $p = 1/2$, puede utilizarse la distribución normal como aproximación de la binomial (bastará con $n > 10$ ya que así, $np > 5$). En este, caso la estadística de la prueba es:

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - n/2}{\frac{1}{2}\sqrt{n}}$$

cuya distribución es aproximadamente normal $N(0,1)$.

EJEMPLO 15.3.

A un grupo de consumidores que consiste de 15 personas se les pidió que califique de 0 a 20, dos marcas A y B de un bien de consumo diario. Las calificaciones obtenidas se dan en al tabla que sigue.

Consumidor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Marca A	10	19	18	14	10	14	09	17	18	20	14	10	16	08	16
Marca B	06	11	08	07	10	16	13	12	11	10	16	13	07	08	15

Utilice la prueba de los signos al nivel de significación 5%, para probar la hipótesis nula de que la mediana de la diferencia en calificación para las dos marcas del bien es cero, contra una alternativa bilateral.

SOLUCION

En la tabla que sigue se indican con los signos "+" y "–"a cada par de valores cuya medición en la primera muestra es mayor y menor respectivamente que la medición en la segunda muestra. Se asigna el 0 a los pares de mediciones iguales. Resultando 13 signos no descartados, y en consecuencia es $n = 13$ el tamaño de la muestra. De los 13 signos, 9 son positivos y 4 son negativos. (Paquete *MCEST*)

Calificaciones de los productos A y B

Jurado	Marca A	Marca B	Signo de la diferencia
1	10	06	+
2	19	11	+
3	18	08	+
4	14	07	+
5	10	10	0
6	14	16	–
7	09	13	–
8	17	12	+
9	18	11	+
10	20	10	+
11	14	16	–
12	10	13	–
13	16	07	+
14	08	08	0
15	16	15	+

Si X es el número de signos negativos de los $n = 13$ signos, entonces, $X \sim B(13, 1/2)$.

Continuando con la prueba, se tiene:

1. *Hipótesis*: $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ contra $H_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$
2. *Nivel de significación*: $\alpha = 0.05$
3. *Estadística*: Variable X binomial $B(n = 13,\ p = 1/2)$.
4. *Cálculos*:
   Para $n = 13, x = 4$ el valor calculado de P es:

$$P = P[X \leq 4 \text{ cuando } p = 1/2] = \sum_{k=0}^{4} C_k^{13}(0.5)^{13} = 0.1334$$

5. *Decisión*. Dado que $2P = 2(0.1334) = 0.2668 > 0.05$, no se rechaza H_0

NOTA. Observe que si utilizamos la aproximación normal:

$$np = 13(0.5) = 6.5 > 5,$$

$$\sigma = \sqrt{npq} = \sqrt{13 \times 0.5 \times 0.5} = 1.8$$

$$P = P[X \leq 4] = P\left[Z \leq \frac{(4+0.5)-6.5}{1.8}\right] = P[Z \leq -1.11] = 0.1334$$

$$P = 0.1334$$

Dado que $2P = 2(0.1334) = 0.2668 > 0.05$, no se rechaza H_0

15.1.5 Dos muestras dependientes: La prueba de Wilcoxon

La prueba de la hipótesis nula que consiste en afirmar que dos muestras correlacionadas provienen de dos poblaciones idénticas utilizando el método del signo, considera solamente el signo de las diferencias entre pares de valores pero ignora las magnitudes, por ejemplo, una diferencia de 12, es tratada en la misma forma que una diferencia de 1.

La **prueba de rango con signo de Wilcoxon** considera los signos e incorpora la magnitud de las diferencias. Por esta razón es considerada la prueba mas potente de todas las pruebas no paramétricas acerca de dos muestras correlacionadas. Además, es una alternativa adecuada a la prueba paramétrica t de dos muestras correlacionadas cuando se verifican los requerimientos de normalidad.

El resumen del procedimiento para la prueba de rango con signo de Wilcoxon para n pares de datos (x_i, y_i) es como sigue:

Ordenar las diferencias $d_i = x_i - y_i$ por valores absolutos (sin tener en cuenta los signos) y descartando las diferencias nulas. **El número de parejas no descartadas es el tamaño *n* de la muestra.**

Se asigna el rango 1 a la mas pequeña d_i, el rango 2 a la siguiente menor, etc. Si hay dos o más diferencias d_i iguales, a todas ellas se les asigna el rango promedio.

Se asigna a cada rango el signo de la diferencia a que corresponde y luego, se calculan las sumas de rangos para diferencias positivas por un lado y para diferencias negativas por otro lado.

La menor de las dos sumas en valor absoluto se designa por T y se conoce como la **estadística *T* de Wilcoxon**.

Para valores pequeños del tamaño n de la muestra y niveles de significación α de 0.01, 0.025, 0.05 para prueba unilateral y 0.02, 0.05, 0.10 para pruebas de dos colas, en la tabla de valores de la T de Wilcoxon se encuentra el valor crítico K. Se rechaza la hipótesis nula si el valor de $T \leq K$. En caso contrario, no se rechaza la hipótesis nula.

Para una prueba bilateral y valores pequeños del tamaño n de la muestra, se tienen los siguientes valores críticos para $\alpha = 0.05$ y $\alpha = 0.01$ respectivamente:

$$T_{\alpha=0.05} = \frac{n^2 - 7n + 10}{5} \quad \text{y}$$

$$T_{\alpha=0.01} = \frac{11n^2}{60} + 5 - 2n$$

Cuando $n \geq 10$ y la hipótesis nula es verdadera, la estadística T tiene una distribución aproximadamente normal con:

$$\text{Media:} \quad \mu_T = \frac{n(n+1)}{4} \quad \text{y}$$

$$\text{Varianza:} \quad \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

En este caso, para determinar la región crítica de la prueba, se puede utilizar la estadística:

$$Z = \frac{T - \mu_T}{\sigma_T}$$

cuya distribución es aproximadamente normal $N(0,1)$.

EJEMPLO 15.4.

Con los datos del ejemplo 15.3. Utilice la prueba de los signos con rango de Wilcoxon al nivel de significación del 5%, para probar la hipótesis nula de que no existe diferencia en las calificaciones promedio para las dos marcas del bien, contra una alternativa bilateral.

SOLUCION

Los datos del ejemplo 15.3 se repiten en la tabla que sigue.

Jurado	Marca A	Marca B	diferencia	Rango	Rango (+)	Rango(–)
1	10	06	4	5.5	5.5	
2	19	11	8	10	10	
3	18	08	10	12.5	12.5	
4	14	07	7	8.5	8.5	
5	10	10	0	-	-	
6	14	16	–2	–25		2.5
7	09	13	–4	–5.5		5.5
8	17	12	5	7	7	
9	18	11	7	8.5	8.5	
10	20	10	10	12.5	12.5	
11	14	16	–2	–2.5		2.5
12	10	13	–3	–4		4
13	16	07	9	11	11	
14	08	08	0	-	-	
15	16	15	1	1	1	
Total					76.5	14.5

En la misma tabla, se asigna el símbolo "+" a cada par de valores cuya medición en la primera muestra es mayor que la medición en la segunda muestra, y se asigna el símbolo "–" en caso contrario.

Si las magnitudes son iguales se asigna 0. Resultando $n = 13$. De los 13 signos, 9 son positivos.

La suma de rangos positivos es 76.5 y la suma de rangos negativos es 14.5, de manera que la menor de las dos sumas es $T = 14.5$ (Paquete *MCEST*)

Continuando con la prueba, se tiene:

1. *Hipótesis*:

   $H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ contra $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$

2. *Nivel de significación*: $\alpha = 0.05$

3. *Región crítica*: Con $n = 13$, $\alpha = 0.05$ y una prueba bilateral, en la tabla de valores de T de Wilcoxon se encuentra el valor crítico 17. Se rechazará H_0 si $T \leq 17$. No se rechazará en caso contrario.

4. *Decisión*. Dado que el valor calculado de T :14.5 es menor que el valor crítico, se rechaza H_0.

NOTA. Observe que si se utiliza el valor crítico bilateral aproximado de T se obtiene:

$$T_{\alpha=0.05} = \frac{n^2 - 7n + 10}{5} = \frac{13^2 - 7(13) + 10}{5} = 17.6$$

y dado que el valor calculado $14.5 \leq 17.6$, se rechaza H_0.

NOTA. Observar que si se utiliza el método de aproximación a la normal se tiene:

$$\mu = \frac{13(13+1)}{4} = 45.5$$

$$\sigma = \sqrt{\frac{13(13+1)(2(13)+1)}{24}} = 14.3$$

$$Z = \frac{14.5 - 45.5}{14.3} = -2.18$$

$P[Z \leq -2.18] = 0.0146$, $2P = 0.0292$, con la que se rechaza H_0.

15.1.6 Dos muestras Independientes: La prueba U de Mann-Whitney.

La prueba de Mann-Whitney denominada también prueba U, se utiliza para probar la hipótesis nula de que **dos muestras aleatorias independientes** provienen de dos poblaciones iguales o de la misma población, cuando no se cumple la suposición de normalidad.

No se requiere hacer suposición alguna acerca de la forma de las distribuciones poblacionales. Pero si se requiere que el nivel de medición de las observaciones de las dos muestras sean al menos de escala ordinal.

La hipótesis nula consiste en suponer que las dos poblaciones son iguales, esto es, $H_0: P_1 = P_2$. La hipótesis alternativa consiste en afirmar que las dos poblaciones son diferentes, esto es, $H_1: P_1 \neq P_2$.

Si las distribuciones de las poblaciones son continuas, la hipótesis nula, también consiste en afirmar que las dos medias poblacionales son iguales, esto es: $H_0: \mu_1 = \mu_2$ contra una alternativa unilateral o bilateral.

Si se verifica la suposición de normalidad, la correspondiente prueba paramétrica *t* para dos muestras independientes es la adecuada.

El procedimiento es el siguiente:

Sea n_1 el número de casos más pequeño de la dos muestras independientes y n_2 el número de casos más grande.

Se combinan las dos muestras en un arreglo ordenado, identificando los valores muestrales de acuerdo con el grupo muestral al que pertenecen.

Se ordenan los valores de menor a mayor, asignando el rango 1 al más pequeño, el rango 2 al siguiente menor valor, etc. Si se encuentran valores iguales, se les asigna el promedio de sus rangos.

La estadística de Mann-Whitney se denota con U. El valor de U se basa en la suma de rangos de cualquiera de las dos muestras, y se define como el menor de los dos valores de U_1 y U_2 que se calculan mediante las siguientes fórmulas:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2$$

donde n_1 = tamaño de la primera muestra

n_2 = tamaño de la segunda muestra

T_1 = Suma de los rangos de la primera muestra.

T_2 = Suma de los rangos de la segunda muestra.

Observar que las T_i son las sumas de rango de Wilcoxon.

Para el nivel de significación α y los tamaños n_1, n_2 de la muestras, en la tabla de valores de la U de Mann-Whitney, se encuentra el valor crítico K. Se rechaza la hipótesis nula H_0 si $U \leq K$. No se rechaza la hipótesis nula en caso contrario.

Cuando $n_i \geq 8$ se admite que la distribución muestral de U es aproximadamente normal, con

$$\text{Media: } \mu_U = \frac{n_1 n_2}{2} \quad \text{y}$$

$$\text{Varianza: } \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Se puede determinar entonces, la significación de un valor observado de U (no importa si es de U_1 o de U_2) por medio de la estadística Z:

$$Z = \frac{U - \mu_U}{\sigma_U}$$

cuya distribución es aproximadamente normal $N(0,1)$.

EJEMPLO 15.5.

Se registraron los siguientes muestras de ingresos familiares mensuales (en dólares) de dos ciudades A y B.

Ingresos A: 850, 600, 700, 910, 870, 800, 700, 1050

Ingresos B: 940, 1000, 1100, 880, 850, 860, 780, 650, 820, 700

Utilice la prueba U de Mann-Whitney al nivel de significación del 1%, para probar la hipótesis nula de que son iguales los ingresos familiares de las dos poblaciones A y B contra una alternativa bilateral.

SOLUCION.

En la tabla que sigue se convierten las 18 observaciones a rangos y se suman los mismos para cada grupo (paquete *MCEST*)

Rangos de los datos

Ingresos A	Ingresos B	Rangos A	Rangos B
850	940	9.5	15
600	1000	1	16
700	1100	4	18
910	880	14	13
870	850	12	9.5
800	860	7	11
700	780	4	6
1050	650	17	2
	820		8
	700		4
$n_1 = 8$	$n_2 = 10$	$w_1 = 68.5$	$w_2 = 102.5$

Continuando con la prueba, se tiene :

1. *Hipótesis*:

   $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$

2. *Nivel de significación*: $\alpha = 0.05$

3. *Estadística*: U de Mann-Whitney (no se supone normalidad)

4. *Región crítica*. Para $n_1 = 8$ y $n_2 = 10$, y $\alpha = 0.05$, en la tabla de valores de U se obtiene el valor crítico igual a 17. La región crítica es $U \leq 17$. Se rechazará H_0 si U, el menor de los valores de U_1 y U_2 es menor o igual que 17. No se rechazará H_0 en caso contrario.

5. *Cálculos*. De la tabla de asignación se rangos se obtiene:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1 = 8 \times 10 + \frac{8 \times (8+1)}{2} - 68.5 = 47.$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2 = 8 \times 10 + \frac{10 \times (10+1)}{2} - 102.5 = 32.5$$

   Para una prueba de dos colas la estadística es la menor de los dos valores de U_1 y U_2, luego, $U = 32.5$

6. *Decisión*. Dado que $u = 32.5 > 17$, no se rechaza H_0.

NOTA. El lector puede verificar que si se supone normalidad, utilizando la prueba t se obtiene $F = 0.093$, significación: 0.764, $t = -0.735$, significación bilateral: 0.473, también se acepta H_0.

15.1.7 *K* muestras independientes: La prueba *H* de Kruskall-Wallis.

La prueba de Kruskall-Wallis, llamada también prueba *H* de Kruskall-Wallis se utiliza para probar la hipótesis nula de que *k* muestras independientes provienen de poblaciones idénticas o de la misma población. Es una generalización de la prueba *U* de Mann-Whitney para el caso de $k > 2$ muestras independientes.

La hipótesis nula H_0 consiste en suponer que las *k* poblaciones son iguales. La hipótesis alternativa H_1 es: las *k* poblaciones no son iguales.

La prueba de Kruskall-Wallis es el equivalente no paramétrico de la prueba de igualdad de *k* medias **del análisis de varianza** para clasificación simple. En su aplicación no se requiere verificar que se cumplan los requisitos necesarios para la validez de las prueba paramétrica correspondiente.

Se requiere que los valores de las muestras aleatorias independientes estén cuando menos en escala ordinal.

El proceso de la prueba es como sigue:

Se combinan las *k* muestras en un arreglo ordenado, identificando los valores muestrales de acuerdo con el grupo muestral al que pertenecen.

Se ordenan los valores de menor a mayor, asignándole el rango 1 al más pequeño, el rango 2 al siguiente más pequeño, etc.. Si se encuentran valores iguales, se les asigna como rango el promedio de sus rangos.

Si la hipótesis nula es verdadera, el promedio de rangos debe ser mas o menos igual para cada uno de los grupos muestrales.

La estadística de la prueba se designa mediante *H*, y se basa en la suma de los rangos de cada una de las diversa muestras aleatorias. Se calcula de la siguiente manera:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

en donde, $n =$ Tamaño de muestra que resulta de sumar los tamaños de cada uno de los grupos muestrales

$R_i =$ Suma de los rangos para *i*-ésima muestra

$n_i =$ Número de observaciones de la *i*-ésima muestra

Si el tamaño de cada una de las muestras es de cuando menos 5, esto es, $n_i \geq 5$ y la hipótesis nula es cierta, la estadística H se distribuye aproximadamente como la distribución χ^2 con $k-1$ grados de libertad donde k es el número de grupos.

Para el nivel de significación α y para $k-1$ grados de libertad en la tabla de valores de chi-cuadrado se encuentra el valor crítico de la prueba: $\chi^2_{1-\alpha,k-1} = c$. Se rechazará H_0 si el valor calculado de $H > c$. En caso contrario no se rechazará la hipótesis nula H_0.

EJEMPLO 15.6.

Los datos que siguen representan los tiempos de operación, en horas, de tres tipos de baterías de teléfonos celulares hasta antes que requiera cargarlas: Utilice la prueba de Kruskall-Wallis, al nivel de significación del 5%, para probar la hipótesis que los tiempos de operación para las tres calculadoras son iguales.

Calculadoras		
A	B	C
24.9	19.8	18.2
26.1	18.2	19.0
22.8	17.5	19.6
21.8	17.5	20.1
25.9	19.2	20.3
	18.5	23.4
	19.0	
	19.5	

SOLUCION

En la tabla que sigue se convierten las 19 observaciones a rangos y se suman los mismos para cada grupo (Paquete *MCEST*)
Continuando con la prueba se tiene

1. *Hipótesis*:

   H_0 :La tres poblaciones son iguales o $H_0 : \mu_1 = \mu_2 = \mu_3$.

   H_1 : Las tres poblaciones no son iguales o no son iguales las tres medias.

2. *Nivel de significación*: $\alpha = 0.05$

3. *Estadística*: $H = \dfrac{12}{n(n+1)} \displaystyle\sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$, que se distribuye aproximadamente como chi-cuadrado con $k-1$ grados de libertad, cuando es verdadera H_0.

4. ***Región crítica.*** Para el nivel de significación $\alpha = 0.05$ y para $k-1=2$ grados de libertad el valor crítico de la prueba es: $\chi^2_{0.95,2} = 5.99$. Se rechazará H_0 si el valor calculado de chi-cuadrado es mayor de 5.99.

5. ***Cálculos.*** De la tabla se asignación se rangos se tiene:
$n_1 = 5$, $n_2 = 8$, $n_3 = 6$, $r_1 = 83$, $r_2 = 46$, $r_3 = 61$, el valor h de la estadística H es:

$$h = \frac{12}{19 \times 20}\left(\frac{(83)^2}{5} + \frac{(46)^2}{8} + \frac{(61)^2}{6}\right) - 3 \times 20 = 11.446$$

6. ***Decisión.*** Dado que $h = 11.446 > 5.99$, se debe rechazar la hipótesis nula. Con el paquete *MCEST* se obtiene $P[H > 11.446] = 0.000$.

Rangos de las duraciones de las operaciones.

Calculadoras		
A	B	C
17	11.0	3.5
19	3.5	6.5
15	1.5	10.0
14	1.5	12.0
18	8.0	13.0
	5.0	16.0
	6.5	
	9.0	
$r_1 = 83$	$r_2 = 46$	$r_3 = 61$

NOTA. El lector puede verificar que si se supone que se verifican los requisitos para el análisis de la varianza paramétrica, utilizando la prueba F se obtiene $F = 304$, gl: 2 y 16, significación: 0.000.

NOTA (Prueba de la mediana para k muestras independientes)

Se puede utilizar la prueba de la mediana para probar la hipótesis nula H_0 de que k $(k \geq 2)$ poblaciones continuas son idénticas con respecto a sus funciones de distribución, contra la alternativa H_1 que es la negación de H_0.

El procedimiento es como sigue:

Hallar la mediana M_e de los n datos que resultan de combinar las k muestras de tamaños $n_1, n_2, \ldots, n_k$.

Comparar los datos de cada muestra con M_e, representando con a_i y con b_i respectivamente a los elementos de la i-ésima muestra que sean mayor que M_e y menor o igual que M_e.

Cuando $n \geq 20$ y si $n_i \geq 5$, se tiene la estadística:

$$\chi^2_{k-1} = \frac{(n-1)}{m_1 \times m_2} \sum_{i=1}^{k} \frac{(n \times a_i - n_i \times m_1)^2}{n \times n_i}$$

cuya distribución se aproxima a una chi-cuadrado con $k-1$ grados de libertad. En la estadística ,

$$m_1 = \sum_{i=1}^{k} a_i \,, \qquad m_2 = \sum_{i=1}^{k} b_i$$

Por ejemplo, al aplicar el método de la mediana a los datos del ejemplo 15.6 se tiene: $n = 19$, $M_e = 19.6$

	Muestra 1	Muestra 2	Muestra 3	Total
$> M_e$	5	1	3	$m_1 = 9$
$\leq M_e$	0	7	3	$m_2 = 10$
Total	$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n = 19$

$$\chi^2_{k-1} = \frac{(19-1)}{9 \times 10} \left[\frac{(19 \times 5 - 5 \times 9)^2}{19 \times 5} + \frac{(19 \times 1 - 8 \times 9)^2}{19 \times 8} + \frac{(19 \times 3 - 6 \times 9)^2}{19 \times 6} \right] = 8.975$$

Para el nivel de significación $\alpha = 0.05$ y 2 grados de libertad el valor crítico de la prueba es: $\chi^2_{0.95,\,2} = 5.99$. Dado que $9.474 > 5.99$, se rechaza la hipótesis nula.

15.1.8 *K* muestras correlacionadas: La prueba *F* de Friedman

La prueba *F* no paramétrica de Friedman es aplicable al análisis de varianza de una clasificación completamente aleatorizado por bloques para el caso en que no se cumplan los requisitos necesarios para la validez de las prueba paramétrica correspondiente. Se requiere que los datos estén al menos en escala ordinal.

Supongamos que se tienen k tratamientos y n bloques de tal manera que los tratamientos deben estar una vez y al azar en cada bloque. Se tienen entonces nk datos organizados en k columnas (muestras, tratamientos o condiciones) y n filas (bloques, grupos o sujetos)

La hipótesis nula H_0, consiste en afirmar que no existe diferencia entre los k tratamientos o que las muestras provienen de la misma población. La hipótesis alternativa bilateral, H_1 consiste en hacer una afirmación contraria a la hipótesis nula.

El proceso de la prueba es como sigue:

Se ordenan los valores de menor a mayor en cada bloque (fila), asignándoles valores ordinales, esto es, asignando el rango 1 al más pequeño, el rango 2 al siguiente, etc.. Si se encuentran valores iguales, se les asigna como rango el promedio de sus rangos.

Se suman los rangos en cada tratamiento (columna)

La estadística de la prueba es:

$$F = \frac{12}{nk(k+1)} \sum_{j=1}^{k} O_j^2 - 3n(k+1)$$

donde k es el número de tratamientos, n es el número de bloques y

$$O_j = \sum_{i=1}^{n} O_{ij} \quad , \quad j = 1,2,\cdots,k$$

es la suma de los valores ordinales obtenida en cada tratamiento j.

Si la hipótesis nula es verdadera, para valores pequeños de n o de k o de ambos, las probabilidades de la distribución muestral de la estadística F están tabulados en una tabla llamada tabla de Friedman, la misma que contiene probabilidades exactas para valores de $k = 3$ y n de 2 a 9, y para $k = 4$ y n de 2 a 4.

Cuando n y k son mayores que tales valores, la distribución muestral de la estadística F, se aproxima a una distribución χ^2 con $k-1$ grados de libertad, donde k es el número tratamientos (Friedman 1937).

Dado el nivel de significación α y para $k-1$ grados de libertad, en la tabla de valores de chi-cuadrado se encuentra el valor crítico de la prueba: $\chi^2_{1-\alpha,k-1} = c$

Se rechazará la hipótesis nula H_0 si el valor calculado de $H > c$. En caso contrario no se rechazará H_0.

EJEMPLO 15.7.

Con el fin de comparar la eficiencia en la preparación preuniversitaria en cuatro academias, se tomaron cinco distintas pruebas a 20 alumnos de cada una de las academias. Los resultados en promedio de los 20 alumnos en cada una de las pruebas se dan a continuación:

Bloques	Academia 1	Academia 2	Academia 3	Academia 4
Prueba 1	17	19	16	12
Prueba 2	18	15	14	13
Prueba 3	16	15	13	14
Prueba 4	16	17	12	11
Prueba 5	19	18	16	12

Utilizando el nivel de significación $\alpha = 0.05$, ¿es compatible con estos resultados experimentales la hipótesis nula H_0: "es igual la eficiencia de la enseñanza en las cuatro academias"?

SOLUCION

Ordenando los resultados por cada prueba (bloques) y sumando los valores ordenados que corresponden a cada academia (tratamiento) se tiene la siguiente tabla de valores ordinales (Paquete *MCEST*)

Valores ordinales

Bloques \	Academia 1	Academia 2	Academia 3	Academia 4
Prueba 1	3	4	2	1
Prueba 2	4	3	2	1
Prueba 3	4	3	1	2
Prueba 4	3	4	2	1
Prueba 5	4	3	2	1
Total	18	17	9	6

La estadística *F* es entonces:

$$F = \frac{12}{5\times 4\times(4+1)}\left(18^2+17^2+9^2+6^2\right)-3\times 5\times(4+1) = 87.6-75 = 12.6$$

Para el nivel de significación $\alpha = 0.05$ y para $k-1=3$ grados de libertad, el valor crítico de la prueba es: $\chi^2_{0.95,3} = 7.81$. Dado que 12.6>7.81, se rechaza la hipótesis nula de igual eficiencia en la enseñanza en las cuatro academias.

15.2. Coeficiente de correlación de rango de Spearman.

El coeficiente de correlación r de Pearson se aplica cuando se supone que hay una relación lineal entre dos variables X e Y cuyas mediciones de las variables son al menos en escala de intervalos. Para probar la hipótesis nula $H_0: \rho = 0$ se supone además distribución normal.

Una medida no paramétrica de la asociación entre dos variables X e Y es el coeficiente de correlación de rangos de Spearman r_s (C. Spearman en 1904). Para n pares de datos observados (x_i, y_i), $i = 1,2,...n$ de (X,Y) el coeficiente de correlación de rangos de Spearman r_s está dado por:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

donde d_i es la diferencia entre los rangos asignados a x_i e y_i.

El coeficiente r_s de Spearman es el coeficiente r de Pearson basado en el rango de los datos y en consecuencia se interpreta en forma similar al coeficiente r de Pearson, esto es: $-1 \leq r_s \leq 1$. Cuando $r_s = +1$ o $r_s = -1$, indica una correlación perfecta entre X e Y. Cuando $r_s = 0$, se concluye que no existe correlación entre las variables X e Y.

Para aplicar el coeficiente r_s de Spearman no se requiere suponer que hay una relación lineal entre X e Y. Tampoco se requiere hacer suposiciones de normalidad respecto a X e Y para probar la hipótesis nula: $H_0: \rho_s = 0$. Se requiere si que las mediciones sean hechas en escala por lo menos ordinal.

Cuando $n \geq 20$ se puede probar la hipótesis nula $H_0: \rho_s = 0$ contra una alternativa adecuada utilizando la estadística:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

que se distribuye aproximadamente como una t-Student con $n-2$ grados de libertad.

Cuando $8 \leq n < 20$ se puede utilizar la prueba t con el coeficiente de Pearson corregido por continuidad $\dot{r}_s$ definida por:

$$\dot{r}_s = 1 - \frac{\sum_{i=1}^{n} d_i}{\frac{n(n^2-1)}{6} + 1}$$

En este caso, la estadística

$$t' = \frac{\dot{r}_s \sqrt{n-2}}{\sqrt{1-\dot{r}_s^2}}$$

tiene aproximadamente distribución t-Student con $n-2$ grados de libertad.

Cuando n pequeña están disponibles las tablas de probabilidades de la distribución exacta de r_s para tomar la decisión de aceptar o rechazar $H_0 : \rho_s = 0$.

EJEMPLO 15.8

La siguiente tabla contiene las calificaciones registradas por 10 alumnos en el examen parcial y final del curso de Estadística Aplicada:

Alumno	1	2	3	4	5	6	7	8	9	10
Examen 1	14	09	15	08	13	16	10	12	07	11
Examen 2	17	13	18	09	16	15	11	12	10	14

a) Calcule el coeficiente de correlación por rangos de Spearman

b) Pruebe la hipótesis nula $H_0 : \rho_s = 0$ contra la alternativa $H_1 : \rho_s > 0$ al nivel de significación $\alpha = 0.5$.

SOLUCION

Se ordenan por rangos las mediciones en cada una de las variables X e Y.

En la tabla que sigue se presentan los rangos de las notas del examen 1 y del examen 2, las diferencias d_i en rango para los diez pares de datos y los cuadrados de las diferencias d_i^2.

Sustituyendo en la formula para r_s se obtiene:

$$r_s = 1 - \frac{6(24)}{10(10^2-1)} = 0.8545$$

El coeficiente de correlación de Spearman corregido por continuidad es

$$\dot{r}_s = 1 - \frac{24}{\frac{10(10^2-1)}{6}+1} = 0.8554$$

Rangos para las notas del examen 1 y examen 2

Alumno	Examen 1	Examen 2	d_i	d_i^2
1	8	9	−1	1
2	3	5	−2	4
3	9	10	−1	1
4	2	1	+1	1
5	7	8	−1	1
6	10	7	+3	9
7	4	3	+1	1
8	6	4	+2	4
9	1	2	−1	1
10	5	6	+1	1
			24	

b)

1. Hipótesis $H_0 : \rho_s = 0$ contra $H_1 : \rho_s > 0$
2. Nivel de significación $\alpha = 0.5$
3. La región crítica es $t_{0.95,8} = 1.860$. Se rechaza H_0 si t' calculada es mayor de 1.860. No se rechaza en caso contrario.
4. La estadística

$$t' = \frac{(0.8554)\sqrt{10-2}}{\sqrt{1-(0.8554)^2}} = 4.67$$

5. Decisión. Se rechaza H_0 y se concluye que existe una correlación significativa entre las notas de los dos exámenes.

NOTA. Con el paquete *MCEST* se obtiene $P[t > 4.67] = 0.001$, tomándose, entonces, la decisión de rechazar la hipótesis nula.

EJERCICIOS

1. En la planta de la firma "MODA" una muestra de los tiempos en minutos que utilizaron 15 operarios para confeccionar una camisa dio los siguientes datos:

   16, 10, 16, 14, 12, 14, 16, 11, 10, 13, 14, 12, 14, 15, 12

   Pruebe al nivel de significación 0.05, la hipótesis en la que se afirma que la mediana de los tiempos empleados es mayor de 13 minutos.

   a) Utilizando la prueba de los signos

   b) Utilizando la correspondiente aproximación a la normal.

   Rp. H_0: Me≤13, a) n=14, x=6 (negativos), P=[X≤6]=0.3953, b) Z=(6.5−7)/1.871=−0.67, P=0.3947, en ambos casos se acepta H_0.

2. El porcentaje de impurezas que un inspector sanitario encontró en una muestra de 15 botellas de chicha morada envasada sin control se dan a continuación:

   2.0, 2.1, 1.8, 1.7, 1.9, 1.4, 1.7, 2.6, 3.0, 1.2, 3.2, 0.9, 1.7, 2.3, 1.7

   Al nivel de significación de 0.05, ¿es posible concluir que la mediana de los porcentajes de impurezas es menor de 2%?

   a) Utilice la prueba de Wilcoxon.

   b) Utilice la correspondiente aproximación a la normal

   Rp. H_0: Me≥2, a) n=14, T=43, K=26, b) Z=(43−52.5)/15.93=−0.596, P=0.2755, en ambos casos se acepta H_0.

3. El número de televisores 29 pulgadas vendidas durante el mes de julio en una muestra aleatoria de 10 tiendas, se reportan como sigue:

   10, 20, 22, 18, 16, 11, 9, 13, 16, 14

   Se desconoce la forma de la distribución y por eso no resulta apropiado utilizar una prueba paramétrica. Al nivel de significación de 0.05, pruebe la hipótesis nula de que la mediana de la población es igual a 12 unidades por tienda contra una alternativa bilateral.

   a) Utilice la prueba del signo.

   b) Aplique la prueba de Wilcoxon.

   Rp. H_0: Me=12, a) n=10, x=3 (negativos), P=[X≤3]=0.172, 2P=0.344, b) T=10, K=8 en ambos casos se acepta H_0.

4. El gerente de producción de la fábrica "MUEBLES" afirma que la mediana del número de sillas metálicas ensambladas por día con la nueva técnica es mayor de 50 unidades. Durante 16 días se ensamblaron el siguiente número de sillas

45, 55, 50, 62, 50, 64, 60, 61, 46, 50, 58, 53, 57, 54, 50, 59

Al nivel de significación de 0.05 pruebe la hipótesis del gerente.

a) Aplicando la prueba de los signos

b) Utilizando la prueba T de Wilcoxon

Rp. H_0: Me≤50, a) n=12, x=2 (negativos), $P=[X\leq 2]=0.0193$, b) T=7, K=17 en ambos casos se rechaza H_0 al nivel 5%.

5. El gerente de producción de la fábrica de confecciones "JHON" quiere probar si la música de fondo suave incrementa la productividad de sus operarios. En una muestra aleatoria de 12 operarios se observaron el número de pantalones producidos por día con y sin música, resultando los siguientes datos:

Operario	1	2	3	4	5	6	7	8	9	10	11	12
Sin música	20	22	25	28	30	32	20	22	25	28	30	33
Con música	22	31	33	30	34	35	25	24	28	32	34	32

Al nivel de significación 0.05 realice una prueba bilateral, para docimar la hipótesis que afirma que son iguales las medias poblacionales de la producción con y sin música

a) Aplicando la prueba del signo

b) Utilizando la correspondiente aproximación a la normal.

c) Aplicando la prueba paramétrica t. Añada la o las condiciones.

Rp. H_0: $\mu_1=\mu_2$, a) n=12, x=1 (positivo), $P=[X\leq 1]=0.0032$, $2P=0.0064$,
b) $Z=(1.5-6)/1.732=-2.598$, $P=0.004$, $2P=0.008$.
c) $t=-4.81$, gl=11, $2P=0.0008$. en todos los casos se rechaza H_0.

6. A un grupo de consumidores se le pide que califique dos marcas de te, de acuerdo con un sistema de evaluación. Las calificaciones se reportan en la siguiente tabla:

Consumidor	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Te 1	46	68	60	58	42	43	40	56	38	58	42	48	45	55
Te 2	36	50	58	40	44	43	29	36	46	48	38	42	40	49

Realice una prueba bilateral al nivel de significación del 5% de la hipótesis nula que afirma que no existe diferencia en las calificaciones medias para las dos marcas de te

a) Utilizando la prueba de Wilcoxon

b) Utilizando la correspondiente aproximación a la normal.
c) Suponiendo población normal y aplicando la prueba t.

Rp. H_0: Me1=Me2 a) n=13, T=8.5, K=17 b) Z=(8.5−14.5)/14.309=−2.586, $2P$=0.0098, c) n=14, t=3.33, gl=13, $2P$=0.0056. en todos los casos se rechaza H_0.

7. Un investigador aplica una misma prueba de aptitud a una muestra de 11 miembros de un grupo de control (X) y a otra muestra de 11 personas asociadas después de un periodo de tensión (Y). Las calificaciones se dan en la tabla que sigue:

X	16	20	17	17	10	14	16	17	13	17	16
Y	06	18	19	10	14	13	02	08	11	12	08

Pruebe la hipótesis de que la mediana de las calificaciones es menor después del periodo de tensión al nivel de significación del 5%

a) Utilizando la prueba de los signos
b) Mediante la prueba de Wilcoxon

Rp. H_0: MeX≤MeY, a) n=9, x=2 (negativos), P=[X≤2]=0.033, b) T=8, K=14 en ambos casos se rechaza H_0 al nivel 5% unilateral.

8. El banco "PRESTA" quiere estudiar el ingreso familiar semanal de dos ciudades A y B. Una muestra aleatoria de A y otra de B han dado los siguientes ingresos en semanales en dólares:

Ciudad A: 150, 154, 143, 160, 95, 70, 83, 97, 135
Ciudad B: 200, 140, 190, 193, 205, 122, 62, 95, 123, 180

Al nivel de significación de 0.05, desarrolle una prueba de hipótesis bilateral para investigar si son iguales los promedios de ingresos en las dos ciudades

a) Usando el método U de Mann-Whitney
b) Utilizando el método paramétrico t. ¿Cuál es el supuesto?

Rp. H_0: $\mu_1=\mu_2$, a) n_1=9, n_2=10, U=29.5, K=20, b) Z=(29.5−45)/12.247=−1.27, $2P$=0.206, c) t=−1.52 (varianza iguales), gl=17, $2P$=0.1428. en todos los casos se acepta H_0.

9. La compañía constructora "PUENTES" estudia las resistencias a la rotura de dos marcas de cable de acero. Dos muestras aleatorias independientes de los cables A y B revelaron las siguientes resistencias a las roturas en libras:

Cable A: 85, 88, 62, 73, 77, 84, 76, 70, 75, 80
Cable B: 67, 68, 64, 66, 74, 78, 65, 70, 72, 73, 65, 70

Al nivel de significación del 5%, pruebe la hipótesis de que la media de la resistencia del cable A es mayor que la media del cable B.

a) Utilizando la prueba U de Mann-Whitney
b) Utilizando la prueba t, si se supone población normal.

Rp. H_0: $\mu_1 \leq \mu_2$ a) $n_1=10$, $n_2=12$, $U=22.5$, $K=34$, $Z=(22.5-60)/15.16=-2.47$, $P=0.007$, b) $t=2.797$ (varianza diferentes), gl=13, $P=0.007$. en todos los casos se rechaza H_0.

10. Para determinar el efecto en el desarrollo psicológico de los escolares que tienen que viajar en trasporte urbano, se tomó una prueba de ansiedad a 12 escolares que usan combis para ir a su colegio y a 15 escolares que van caminando de. Las calificaciones registradas son las siguientes:

En combi: 14, 18, 16, 08, 16, 14, 13, 12, 09, 18, 07, 15
Caminando: 18, 15, 09, 18, 10, 17, 15, 15, 06, 15, 12, 15, 19, 16, 11

Al nivel de significación de 0.05 y con una prueba bilateral, ¿se puede concluir que el uso de combis para ir al colegio no aumenta la ansiedad en los escolares?.

a) Mediante la prueba de la U de Mann-Whitney
b) Utilizando la correspondiente aproximación a la normal.
c) Suponiendo población normal, aplique la prueba t.

Rp. H_0: $\mu_1=\mu_2$ a) $n_1=12$, $n_2=15$ $U=77.5$, $K=49$, b) $Z=(77.5-90)/20.494=0.61$, $2P=0.541$, c) $t=-0.51$ (varianza iguales), gl=25, $2P=0.6114$. en todos los casos se acepta H_0.

11. Para evaluar y comparar dos métodos de capacitación industrial, un director de capacitación asigna al azar a cada uno de dos métodos. Debido a la deserción normal, 12 aprendices terminaron el curso mediante el método 1 y 10 terminaron llevando el método 2. A los dos grupos se les aplicó el mismo examen para evaluar el aprendizaje, resultando las siguientes calificaciones.

Método 1: 75, 95, 87, 69, 81, 72, 84, 76. 76, 61, 68, 65
Método 2: 83, 75, 87, 79, 62, 84, 85, 82, 73, 91

Al nivel de significación del 5%, pruebe la hipótesis nula de que la mediana del método1 es igual a la mediana del método 2, contra una prueba bilateral

a) Utilizando la prueba U de Mann-Whitney
b) Utilizando la prueba de la mediana

Rp. H_0: $\mu_1 \leq \mu_2$ a) $n_1=10$, $n_2=12$, $U=22.5$, $K=34$, $Z=(22.5-60)/15.16=-2.47$, $P=0.007$, b) $t=2.797$ (varianza diferentes), gl=13, $P=0.007$. en todos los casos se rechaza H_0.

12. La compañía "MICA" produce camisas para caballeros en sus dos talleres de Lima. El Jefe de control de la calidad afirma que el número de camisas con defectos del taller A no es mayor que el número de camisas con defectos del taller B. Dos muestras aleatorias del número de camisas con defectos de la producción de 11 días dieron los siguientes resultados:

Taller A: 14, 9, 13, 10, 12, 11, 14, 13, 10, 12, 15
Taller B: 9, 11, 7, 6, 5, 8, 9, 10, 9, 6, 12

Al nivel de significación del 5%, investigue si tiene razón el jefe de control de la calidad
a) Utilizando la prueba de la mediana
b) Utilizando la prueba paramétrica *t*. Suponga distribución normal.

Rp. H_0: $\mu_1 \leq \mu_2$ a) Me=10, >Me=10, ≤Me=12, n_1=11, n_2=11, H=6.6, P=0.001, b) t=4.22. (varianzas iguales), gl=20, P=0.003. en todos los casos se rechaza H_0.

13. En el departamento de control de calidad de la empresa "EQUIP" se desea comparar el tiempo que se requiere para diagnosticar fallas de equipo utilizando dos sistemas alternativos. Se asigna al azar una muestra de 26 fallas de equipo para diagnosticarlas mediante dos sistemas. El primer sistema se utiliza para diagnosticar 12 fallas y el segundo para diagnosticar 14 fallas. El tiempo total en minutos que se utilizaron para diagnosticar las fallas son las siguientes:

Sistema 1: 20, 24, 37, 11, 26, 9, 28, 40, 21, 29, 25, 38
Sistema 2: 20, 35, 38, 54, 47, 26, 19, 35, 33, 38, 42, 30, 60, 36

Al nivel de significación del 5%, pruebe la hipótesis de que la mediana del sistema1 es menor que la mediana del sistema2
a) Utilizando la prueba U de Mann-Whitney
b) Utilizando la correspondiente aproximación normal

Rp. H_0: Me1≥Me2 a) n_1=12, n_2=14, U=43, K=45, b) Z=(43−84)/19.462=−2.1, P=0.017, en ambos casos se rechaza H_0.

14. El número de horas de vida útil de muestras de dos tipos de baterías A y B para calculadoras científicas se registraron como sigue:

Batería A: 4.4, 4.6, 6.1, 3.5, 4.2, 4.0, 5.1, 4.9
Batería B: 2.9, 3.7, 3.2, 3.5, 3.0, 3.8, 3.6, 4.2

Utilice la prueba de la suma de rangos de Wilcoxon (o U de Man-Whitney) con α=0.05 para verificar que la media de la marca A es mayor que la media de la marca B.

Rp. H_0: $\mu_1 \leq \mu_2$ a) n_1=8, n_2=8, U=6, K=13, se rechaza H_0.

15. Se aplicó una prueba de aptitud a una muestra de mujeres y a otra de hombres resultando las siguientes calificaciones:

Mujeres : 13, 12, 12, 10, 10, 10, 10, 9, 8, 8, 7, 7, 7, 7, 7, 6
Hombres: 17, 16, 15, 15, 15, 14, 14, 14, 13, 13, 13, 12, 12, 12, 12,
11, 11, 10, 10, 8, 8, 6

Al nivel de significación del 5%, pruebe la hipótesis de que la media de aptitud es la misma para ambos sexos. Utilice una prueba bilateral y la correspondiente aproximación normal de la estadística U de Mann-Whitney

Rp. H_0: $\mu_1=\mu_2$, $n_1=16$, $n_2=22$, $U=59$, $Z=(59-176)/33.823=-3.46$, $P=0.0007$, $2P=0.0014$, se rechaza H_0.

16. En la tabla que sigue figuran el número de quejas observadas por mes en tres tiendas de la cadena TO&TO:

Sucursal A: 14, 18, 16, 4, 14
Sucursal B: 15, 20, 16, 15, 10, 9, 8, 6
Sucursal C: 15, 16, 9, 10, 6, 5, 4, 7, 11

Utilice la prueba H de Kruskal-Wallis, al nivel de significación de 0.01, para probar si el número de quejas para las tres sucursales, son iguales.

Rp. H_0: $\mu_1=\mu_2=\mu_3$ $H=2.99$, gl=2, $P=0.3167$, se acepta H_0.

17. Se asignaron en forma aleatoria 18 participantes de un programa técnico a 3 tipos distintos de métodos de instrucción en diseño auxiliado por computadora. Las calificaciones obtenidas son las siguientes:

Método 1: 88, 77, 79, 68, 82, 90
Método 2: 91, 77, 89, 83, 88, 87
Método 3: 77, 63, 68, 66, 76, 74

Suponga que las tres poblaciones no tienen distribución normal. Aplique el método de la mediana para probar la hipótesis nula de que las tres poblaciones tienen la misma mediana. Hágalo al nivel de significación 0.05

Rp. H_0: Me1=Me2=Me3, Me=78, >Me=9, ≤Me=9, $n_1=n_2=n_3=6$, $H=9.3$, gl=2, $P=0.009$, se rechaza H_0.

18. Para investigar los rendimientos de tres marcas de gasolina A, B y C en términos de los kilómetros recorridos por litro, se obtuvieron los siguientes datos muestrales:

Gasolina A: 6.4, 5.8, 7.5, 10.0, 9.0, 8.9, 7.1
Gasolina B: 8.3, 10.3, 9.5, 8.8, 10.7, 11.4, 10.2, 10.9
Gasolina C: 11.5, 9.6, 11.8, 12.1, 10.4, 12.3, 13.8, 12.7, 13.8

¿Qué conclusiones puede usted sacar con la prueba H?.

Rp. H_0: $\mu_1=\mu_2=\mu_3$, $H=15.305$, gl=2, $P=0.0004$, se rechaza H_0.

19. Una muestra aleatoria de las ventas diarias en miles de dólares registradas en cada uno de tres hipermercados es:

Mercado A: 4.3, 4.2, 4.9, 5.0, 5.6, 5.2, 4.8
Mercado B: 5.3, 5.7, 5.4, 6.2, 6.5, 5.8, 5.3, 5.1, 6.4
Mercado C: 6.1, 6.4, 6.7, 7.3, 6.6, 6.8, 7.4, 7.9, 7.5, 7.8, 6.9

Con la prueba de Kruskal- Wallis, al nivel de significación de 0.01, ¿se puede inferir que las ventas medias diarias de los tres hipermercados son iguales?

Rp. H_0: $\mu_1=\mu_2=\mu_3$, H=20.245, gl=2, P=0.0000, se rechaza H_0.

20. Tres profesores enseñan a tres secciones del mismo curso de matemáticas. Una muestra aleatoria de las calificaciones registradas en cada sección es:

Sección A: 13, 14, 12, 08, 10, 09, 07, 08, 06, 15
Sección B: 15, 16, 14, 17, 18, 16, 15, 17, 16, 18, 19
Sección C: 12, 11, 1 4, 10, 11, 13, 14, 12, 12, 14, 11, 14, 12

Utilice la prueba de Kruskal-Wallis, al nivel de significación de 0.05, para determinar si las distribuciones de las calificaciones otorgadas por los tres profesores difieren en forma significativa.

Rp. H_0: $\mu_1=\mu_2=\mu_3$, a) H=21.33, gl=2, P=0.0000, se rechaza H_0.

21. Con el objeto de verificar el contenido de alquitrán, se probaron muestras aleatorias de cuatro marcas de cigarros. Las siguientes cifras en miligramos corresponden al alquitrán encontrado en los 16 cigarrillos probados

Marca A	Marca B	Marca C	Marca D
12	12	15	19
08	17	13	21
09	13	12	18
11	15	13	22

Utilice la prueba de Kruskal-Wallis, al nivel de significación de 0.05, para probar si existe una diferencia significativa en el contenido de alquitrán entre las cuatro marcas de cigarros.

Rp. H_0: $\mu_1=\mu_2=\mu_3=\mu_4$, a) H=12.28, gl=3, P=0.00648, se rechaza H_0.

22. Un analista estudia el número de minutos de publicidad en tres canales de TV. Una muestra aleatoria de los tiempos de publicidad registradas en cada CTV es:

Canal A: 13, 14, 12, 08, 10, 09, 07, 08, 06, 15
Canal B: 15, 16, 14, 17, 18, 16, 15, 17, 16, 18, 19
Canal C: 12, 11, 14, 10, 11, 13, 14, 12, 12, 14, 11, 14, 12

Utilice la prueba de Kruskal-Wallis, al nivel de significación de 0.05, para investigar si las distribuciones de los tiempos de publicidad de los tres canales difieren en forma significativa.

Rp. H_0: $\mu_1=\mu_2=\mu_3$, a) H=21.33, gl=2, P=0.0000, se rechaza H_0.

23. En un experimento con 6 tratamientos, en el que todos tienen 5 repeticiones excepto uno que tiene 6, se obtuvieron las ordenes de mérito que se dan en la tabla que sigue. Al nivel de significación de 0.05, ¿se puede concluir que son iguales las medias de las poblaciones de donde provienen las muestras.

Desarrolle la prueba con el método

a) De Kruskal-Wallis,

b) De la mediana

Tratamientos					
A	B	C	D	E	F
15	20	18	21	5	22
17	22	11	23	4	19
16	23	14	24	7	14
18	26	17	19	8	13
16	25	10	22	9	12
14	27	11	20	8	16
15					

Rp. H_0: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5=\mu_6$, a) H=29.329, gl=5, P=0.0000, se rechaza H_0. b) Me=16, >Me=18, ≤Me=19, H=20.05, gl=5, P=0.00122.

24. Un analista estudia las calificaciones de tres grupos de sujetos bajo cuatro condiciones. Cada grupo contiene 4 sujetos que son asignados a cada una de las cuatro condiciones. Los calificaciones se dan en la siguiente tabla:

	Condición 1	Condición 2	Condición 3	Condición 4
Grupo A	19	14	11	17
Grupo B	16	15	12	18
Grupo C	19	11	12	16

Utilice el método del análisis de varianza por rangos F de Friedman al nivel de significación de 0.05, para investigar si son iguales las cuatro condiciones.

Rp. H_0: $\mu_1=\mu_2=\mu_3=\mu_4$ o las cuatro condiciones son iguales, F=7.7, gl=3, P=0.0526, se acepta H_0.

25. Cuatro gerentes de producción evaluaron cinco diseños de un nuevo juguete educativo en una escala de 0 a 20. Los resultados obtenidos fueron los siguientes:

Evaluador	Diseño 1	Diseño 2	Diseño 3	Diseño 4	Diseño 5
E1	15	16	14	08	16
E2	14	12	10	06	14
E3	16	18	11	13	15
E4	18	14	11	09	12

Utilice el método del análisis de varianza por rangos F de Friedman al nivel de significación de 0.05, para probar la hipótesis nula que los 5 diseños no difieren entre si.

Rp. H_0: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_4$ o los 5 diseños son iguales, F=12.35, gl=4, P=0.0149, se rechaza H_0.

26. Los datos que se dan en la tabla que sigue son los costos y las ventas mensuales, en miles de dólares para 14 farmacias:

a) Calcule el coeficiente de correlación por rango r_s.

b) ¿Diría usted que el coeficiente es significativo al nivel $\alpha = 5\%$?

c) Calcule el coeficiente de correlación r de Pearson,

Farmacias	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Costos	10	09	13	12	11	19	20	14	21	17	18	15	16	22
Ventas	18	14	19	13	15	32	31	17	28	21	22	19	20	33

d) Suponga una población bivariante normal. ¿Es significativo el coeficiente r de Pearson al nivel $\alpha = 5\%$?

Rp. a) r_s=0.927, b) H_0: ρ_s=0, t=8.531, gl=12, signific. a dos colas: **0.000 se rechaza H_0.** c) r=0.901, b) H_0: ρ=0, t=7.188, gl=12, signific. a dos colas: **0.000 se rechaza H_0.**

27. En una competencia de habilidades musicales dos jueces calificaron a 10 candidatos. Las calificaciones rangueadas se muestran en la tabla que sigue:

Candidato	A	B	C	D	E	F	G	H	I	J
Juez 1	8	9	6	2	1	4	5	7	3	10
Juez 2	7	10	8	5	3	2	4	6	1	9

a) Calcule el coeficiente de correlación por rango r_s.

b) ¿Diría usted que el coeficiente es significativo al nivel $\alpha = 5\%$?

Rp. a) r_s=0.819, b) H_0: ρ_s=0, t=4.041, gl=8, signific. a dos colas: **0.000, se rechaza H_0.**

28. Deduzca la fórmula de correlación gradual de Spearman

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Rp. Los n valores ordenados de X: son :1,2,...,n, $\bar{x} = n(n+1)/2n = (n+1)/2$, $s_x^2 = n(n+1)(2n+1)/6\times 2 - ((n+1)/2)^2 = (n^2-1)/12$ Análogamente para Y. Por otro lado $s_d^2 = s_d^2 + s_y^2 - 2r_s s_x s_y$, $r_s = \dfrac{s_x^2 + s_y^2 - s_d^2}{2s_x s_y}$ Además, $\bar{d} = 0$, $s_d^2 = (\sum d^2)/n$.

Sustituyendo, se obtiene la fórmula.

Capítulo 16

INTRODUCCION A LA TEORIA DE LA DECISION ESTADISTICA.

16.1 Introducción

Dada una situación en la que se dispone de dos o **más cursos** de acción alternativas o estrategias o simplemente *acciones,* cada uno de los cuales puede llevar a un conjunto de resultados mutuamente excluyentes asociados con ciertas probabilidades, la teoría de la decisión estadística consiste en elegir alguna de estas acciones como la mejor.

Como es obvio, cuando se dispone de una sola acción, no se requiere decisión, por que se debe seguir esa única acción como la mejor decisión. Por lo tanto, se requiere decisión estadística en una situación en la que se dispone por lo menos de dos acciones y en la que sólo se puede seguir una de esas acciones como la mejor.

El proceso de la decisión estadística, luego de planteado el problema, consiste en elaborar una lista de posibles acciones y de posibles eventos de decisión mutuamente excluyentes, asignar probabilidades a cada uno de los eventos cuando estos estén disponibles y luego proceder a tomar la decisión eligiendo una sola alternativa como la mejor acción.

Las **acciones** son las alternativas de acción o estrategias disponibles para la toma de decisiones.

Los **eventos** son las ocurrencias que determinan el nivel del éxito de una acción determinada. Estas están fuera del control de quien toma las decisiones..

Los posibles *eventos* son considerados por un lado como ***eventos mutuamente excluyentes*** y por otro lado como *valores numéricos de una variable aleatoria* discreta o continua. En este capítulo sólo se hará una breve introducción a los eventos discretos. En los ejercicios de este capitulo se incluyen algunos problemas con variables aleatorias continuas.

La teoría de la decisión estadística consiste de una conjunto de técnicas para elegir la mejor acción o la acción más óptima.

El *primer paso*, consiste pues, en expresar el problema de la decisión estadística, confeccionando una lista de todos las posibles *acciones* a tomar y una lista de todos los posibles *eventos*.

EJEMPLO 16.1.

Un fabricante produce cierto bien de consumo a un costo de $5 la unidad y lo vende a un precio de $10 cada uno. Este bien de consumo, produce el fin de semana y vende durante la semana siguiente. El producto es tal que si no lo vende durante la semana, el fabricante lo deshecha perdiendo $2.5 por unidad además del costo.

De acuerdo con la experiencia, la demanda semanal de este producto varía entre 12 y 15 unidades inclusive.

Se trata de decidir cuántas unidades del bien de consumo se debería producir para las ventas de cada semana.

Identifique los eventos las acciones posibles

SOLUCION

En este ejemplo, obviamente, la demanda semanal consta de 4 valores posibles. Estas demandas son los eventos posibles:

E_1: vender 12 unidades
E_2: vender 13 unidades
E_3: vender 14 unidades
E_4: vender 15 unidades

Se tienen cuatro acciones posibles a tomar:

A_1: Producir 12 unidades
A_2 : Producir 13 unidades
A_3 : Producir 14 unidades
A_4 : Producir 15 unidades

Asignación de probabilidades.

El *segundo paso* es asignar probabilidades a los posibles eventos. Un método es utilizar las frecuencias relativas de los datos como probabilidades *iniciales* o *probabilidades a priori*.

Como los eventos son mutuamente excluyentes, la suma de las probabilidades de los eventos debe ser igual a uno.

EJEMPLO 16.2.

Continuando con el ejemplo 16.1, supongamos que durante 1,000 semanas se han registrado los siguientes niveles de ventas: 300 semanas se demandaron 12 unidades, 400 semanas se demandaron 13 unidades, 200 semanas se demandaron 14 unidades, 100 semanas se demandaron 15 unidades, asignar probabilidades a los eventos

SOLUCION

La asignación de *probabilidades a priori* es como sigue: Al evento E_1 se le asigna la probabilidad 0.3, al evento E_2, la probabilidad 0.4, al evento E_3 la probabilidad 0.2, y al evento E_4, la probabilidad 0.1.

16.2 Tabla de pagos

El *tercer paso* en la teoría de la decisión estadística, es construir una *tabla de pagos* que contenga la lista de las acciones alternativas, los posibles eventos y los **pagos** como se indica en la tabla 16.1.

El pago se define como la utilidad neta es decir ventas menos costos.

Los *pagos* X_{ij} de la tabla de pagos son valores que consisten de pérdidas o ganancias que dependen del evento E_i y de la acción A_j.

Un pago es un *valor condicional* en el sentido de que el resultado económico que se experimenta depende de la acción y del evento que ha ocurrido.

La estructura de una tabla de pagos contiene todos los valores condicionales de todas las combinaciones posibles de acciones y eventos de decisión. También contiene la probabilidad de ocurrencia para cada uno de los eventos mutuamente excluyentes, como se indica en el siguiente cuadro.

Tabla 16.1, *Estructura general de una tabla de pagos*

Eventos	Probabilidad	Acciones			
		A_1	A_2	...	A_n
E_1	P_1	X_{11}	X_{12}	...	X_{1n}
E_2	P_2	X_{21}	X_{22}	...	X_{2n}
E_3	P_3	X_{31}	X_{32}	...	X_{3n}
...	...	...	...	...	...
E_k	P_k	X_{k1}	X_{k2}	...	X_{kn}

EJEMPLO 16.3

Continuando con el problema 16.1 determinar la tabla de pagos que se generan a partir del enunciado

SOLUCION

Sea D la demanda semanal del producto cuyos valores son 12, 13, 14, 15 y sea Q la cantidad de unidades que se producen: 12, 13, 14, 15.

El *pago* o *ganancia* o *utilidad* es:

$$pago = ventas - costos$$

Si se producen $Q = 12$ unidades, los pagos para las cuatro demandas respectivas D=12, 13, 14, 15 son iguales a: $12\times10 - 12\times5 = 12\times5 = 60$

Si se producen $Q = 13$ unidades; entonces,

Si $D = 12$, el pago es: $12\times10 - 13\times5 - 1\times2.5 = 52.5$

Si $D = 13$, 14 o 15, el pago es: $13\times10 - 13\times5 = 65$

Si se producen $Q = 14$ unidades; entonces,

Si $D = 12$, el pago es: $12\times10 - 14\times5 - 2\times2.5 = 45$

Si $D = 13$, el pago es: $13\times10 - 14\times5 - 1\times2.5 = 57.5$

Si $D = 14$ o 15, el pago es: $14\times10 - 14\times5 = 70$

Si se producen $Q = 15$ unidades; entonces,

Si $D = 12$, el pago es: $12\times10 - 15\times5 - 3\times2.5 = 37.5$

Si $D = 13$, el pago es: $13\times10 - 15\times5 - 2\times2.5 = 50$

Si $D = 14$, el pago es: $14\times10 - 15\times5 - 1\times2.5 = 62.5$

Si $D = 15$, el pago es: $15\times10 - 15\times5 = 75$

Estos resultados se resumen en la tabla de pagos 16.2

Tabla 16.2. *Tabla de pagos con asignación de probabilidades para una decisión comercial*

Demanda del mercado	Probabilidad	Cantidad de producción			
		A_1:12	A_2:13	A_3:14	A_4:15
E_1:12	0.3	60	52.5	45	37.5
E_2:13	0.4	60	65	57.5	50
E_3:14	0.2	60	65	70	62.5
E_3:15	0.1	60	65	70	75

NOTA. Los doce pagos X_{ij} asociados a cada acción (A_1:12, A_2:13, A_3:14, A_4:15) dado que ha ocurrido un evento especifico (E_1:12, E_2:13, E_3:14, E_4:15) se pueden expresar mediante la función que sigue, llamada ***función utilidad.***

$$Pagos\ X_{ij} = \begin{cases} 10D - 5Q - 2.5(Q - D), & \text{si } D < Q \\ 10Q - 5Q, & \text{si } D \geq Q \end{cases}$$

16.3. Toma de decisiones

Es el *cuarto paso* en el proceso de la toma de decisiones estadísticas. Existen varios criterios que se pueden utilizar para determinar la mejor acción.

1. Criterio basado sólo en probabilidades

El criterio basado sólo en probabilidades sin tener en cuenta las consecuencias económicas, consiste en identificar el evento que tenga *máxima probabilidad* de ocurrencia y elegir la acción que corresponda a ese evento como la más optima.

Continuando con el ejemplo 16.1, el evento con máxima probabilidad es E_2:13 para el cual la probabilidad 0.4. Con base al criterio de máxima probabilidad la mejor acción sería A_2: elaborar 13 unidades del producto cada fin de semana para su venta en la siguiente semana.

2. Criterio basado sólo en las consecuencias económicas.

Para la toma de decisiones basado sólo en las consecuencias económicas sin tener en cuenta las probabilidades se utilizan tres criterios: El criterio maximin, el criterio maximax y criterio del arrepentimiento minimax .

2a. El criterio maximín

Consiste en determinar el valor mínimo que resulta de cada acción a tomar en la tabla de pagos y elegir como la mejor acción aquella cuya resultante es mayor.

Es una decisión muy conservadora pues el que toma este tipo de decisiones se preocupa por *lo peor que puede pasar* con respecto a cada acción.

EJEMPLO 16.4.
Continuando con el ejemplo 16.1, determine la mejor acción a tomar utilizando el criterio maximin.

SOLUCION
En la última fila de la tabla 16.3 se enlista el pago mínimo de cada una de las acciones: A_i, i=1, 2, 3, 4.

El mayor de estos cuatro valores **mínimos** (el **máximo** de los mínimos) es 60.
Por lo tanto, desde el punto de vista del criterio maximin la mejor acción a tomar es A_1: producir 12 unidades cada **fin de semana.**

Tabla 16.3. *Criterio maximin*

Demanda del mercado	Cantidad de producción			
	A_1:12	A_2:13	A_3:14	A_4:15
E_1:12	60	52.5	45	37.5
E_2:13	60	65	57.5	50
E_3:14	60	65	70	62.5
E_3:15	60	65	70	75
Mínimos	60	52.5	45	37.5

2b. El criterio maximax.

Consiste en determinar el valor máximo que resulta de cada acción a tomar en la tabla de pagos y elegir como la mejor acción aquella cuya resultante es mayor.

Es una decisión opuesta al criterio maximin pues el que toma las decisiones se preocupa por *lo mejor que puede suceder* con respecto a cada acción.

EJEMPLO 16.5.
Continuando con el ejemplo 16.1, determine la mejor acción a tomar utilizando el criterio maximax.

SOLUCION
En el último renglón de la tabla 16.4 se enlista el pago máximo de cada acción A_i, i=1, 2, 3, 4.

Tabla 16.4. *Criterio máximax*

Demanda del mercado	Cantidad de producción			
	A_1:12	A_2:13	A_3:14	A_4:15
E_1:12	60	52.5	45	37.5
E_2:13	60	65	57.5	50
E_3:14	60	65	70	62.5
E_3:15	60	65	70	75
Máximos	60	65	70	75

El mayor de estos cuatro valores máximos (el máximo de los máximos) es 75.

Por lo tanto, desde el punto de vista del criterio maximax la mejor acción es A_4: producir 15 unidades.

2c. El criterio del arrepentimiento minimax o de pérdida de oportunidad condicional.

Este método de decisión se basa en los *arrepentimientos*.

Un *arrepentimiento o pérdida de oportunidad condicional* se define como la cantidad de pago perdido al no tomar la acción del pago mas alto para cada evento posible.

Por el criterio del arrepentimiento minimax se identifica como la mejor acción aquella para la cual el arrepentimiento máximo posible es menor.

EJEMPLO 16.6.
Continuando con el ejemplo 16.1, determine la mejor acción a tomar utilizando el criterio del arrepentimiento mínimax.

SOLUCION

Si la demanda es 12 unidades, la acción optima es producir 12 unidades, pues se obtendría una ganancia de $60. Si la decisión es producir 13 unidades, la ganancia sería sólo de $52.5 y se tendría una pérdida de oportunidad de $7.5 (=60 − 52.5). Si la decisión es producir 14 unidades, la ganancia sería sólo de $45 y se tendría una pérdida de oportunidad de $15 (=60 − 45). Y si la decisión es producir 15 unidades, la ganancia sería sólo de $37.5 y se tendría una pérdida de oportunidad de $22.5 (=60–37.5).

Tabla 16.5. *Tabla de pérdidas de oportunidad y aplicación del criterio del arrepentimiento minimax*

Demanda del mercado	Cantidad de producción			
	A_1:12	A_2:13	A_3:14	A_4:15
E_1:12	60 − 60 0	60 − 52.5 7.5	60 − 45 15	60 −37.5 22.5
E_2:13	65 −60 5	65 − 65 0	65 − 57.5 7.5	65 − 50 15
E_3:14	70 − 60 10	70 − 65 5	70 − 70 0	70 − 62.5 7.5
E_4:15	75 − 60 15	75 − 65 10	75 − 70 5	75 − 75 0
Máximo	15	10	15	22.5

Con cálculos similares para las demandas de 13, 14 y 15 unidades se obtienen las demás pérdidas de oportunidad. Todas estas pérdidas de oportunidad se indican en la tabla 16.5

Note que la perdida de oportunidad de cualquier acción óptima es cero.

En la última fila de la tabla 16.5 se enlista el máximo arrepentimiento que puede ocurrir de acuerdo con cada acción A_i . El menor de estos máximos es 10. Entonces, la acción optima desde el punto de vista del arrepentimiento mínimax es A_2: producir 13 unidades.

3. Criterio basado en probabilidades y Consecuencias económicas.

En este rubro hay dos procedimientos que se explican a continuación.

3a. Criterio del pago esperado (PE)

El procedimiento consiste primero en encontrar los valores esperados de los pagos o utilidades esperadas en cada posible acción, luego, por el criterio del pago esperado (PE) llamado también **criterio Bayesiano,** la mejor acción es aquella que tiene el mayor resultado económico **esperado.**

EJEMPLO 16.7.

Continuando con el ejemplo 16.1, determine la mejor acción a tomar utilizando el criterio del pago esperado.

SOLUCION

En la última fila de la tabla 16.6 se enlistan los pagos esperados correspondientes a cada acción A_i, i=1,2,3,4.Por el criterio del pago esperado se elige como acción óptima la acción que tiene mayor pago esperado (PE), esto es, la acción A_2 producir 13 unidades.

Tabla 16.6. *Tabla de pagos para la aplicación del criterio del pago esperado (PE)*

Demanda del mercado	Probabilidad	Cantidad de producción			
		A_1:12	A_2:13	A_3:14	A_4:15
E_1:12	0.3	60	52.5	45	37.5
E_2:13	0.4	60	65	57.5	50
E_3:14	0.2	60	65	70	62.5
E_4:15	0.1	60	65	70	75
Pago esperado (PE)		60	61.25	57.5	51.25

3b. Criterio de la pérdida de oportunidad esperada (*POE*)

La pérdida de la oportunidad esperada (*POE*) es un criterio de decisión alternativo que nos lleva a la misma decisión que el criterio del pago esperado. Según este criterio la mejor acción es aquella que minimiza las pérdidas de oportunidad esperada.

Los cálculos de las *POE* son similares a los de los *PE* excepto que se usa pérdidas de oportunidad *PO* en vez de pagos.

EJEMPLO 16.8.
Continuando con el ejemplo16.1, determine la mejor acción a tomar utilizando el criterio de la pérdida de oportunidad esperada.

SOLUCION

En la última fila de la tabla 16.7 se dan las pérdidas de oportunidad esperada para cada acción. La menor de estas pérdidas de oportunidad esperadas es igual a 4.25, Por el criterio *POE*, la mejor acción es A_2, producir 13 unidades.

Tabla 16.7. *Tabla de pérdidas de oportunidad y Cálculo de las pérdidas de oportunidad esperada (POE).*

Demanda del mercado	Probabilidad	Cantidad de producción			
		A_1:12	A_2:13	A_3:14	A_4:15
E_1:12	0.3	0	7.5	15	22.5
E_2:13	0.4	5	0	7.5	15
E_3:14	0.2	10	5	0	7.5
E_4:15	0.1	15	10	5	0
Pérdida de oportunidad esperada (POP)		5.5	4.25	8	14.25

16.4. Arboles de decisión

Un árbol de decisión es una gráfica o diagrama que corresponde a un proceso de decisiones secuenciales. Contiene ***puntos de decisión***, que son los puntos secuenciales donde debe tomarse una decisión, y contiene ***eventos aleatorios*** que son los puntos secuenciales donde ocurre algún evento probabilístico.

El análisis de árboles de decisión es un método que se utiliza para identificar la mejor acción inicial y las mejores acciones subsiguientes. El criterio de decisión es el criterio Bayesiano del pago esperado o utilidad esperada.

EJEMPLO 16.9
Un fabricante tiene un proyecto para fabricar un producto nuevo y debe decidir si lo realiza o no. El estima un costo de $50,000 para realizar el proyecto. Si el proyecto no tuviera el éxito se termina con el mismo. Si el proyecto tuviera éxito, entonces, el fabricante debe decidir por una producción alta o baja. Si la producción fuera alta, entonces, las ventas se estiman en $150,000 si la demanda fuera alta y en $ 40,000 si la demanda fuera baja. Así mismo, si la producción fuera baja, las ventas se estiman en $70,000 si la demanda fuera alta y en $ 35,000 si la demanda fuera baja. Por otro lado, se estima una probabilidad de éxito igual a 0.8 y de fracaso igual a 0.2 si se realiza el proyecto. Así mismo, se estima una probabilidad de demanda alta igual a 0.35 y la de demanda baja igual 0.65.
Elabore un diagrama de árbol para este problema.

SOLUCION

Utilizando las notaciones:

R=Realizar el proyecto,	E=Éxito,	PA=Producción alta
NR= No realizar el proyecto,	F=Fracaso,	PB=Producción baja
DA=Demanda Alta,	DB=Demanda baja	G=Ganacia

El diagrama de árbol de decisiones de este problema es la figura 16.1

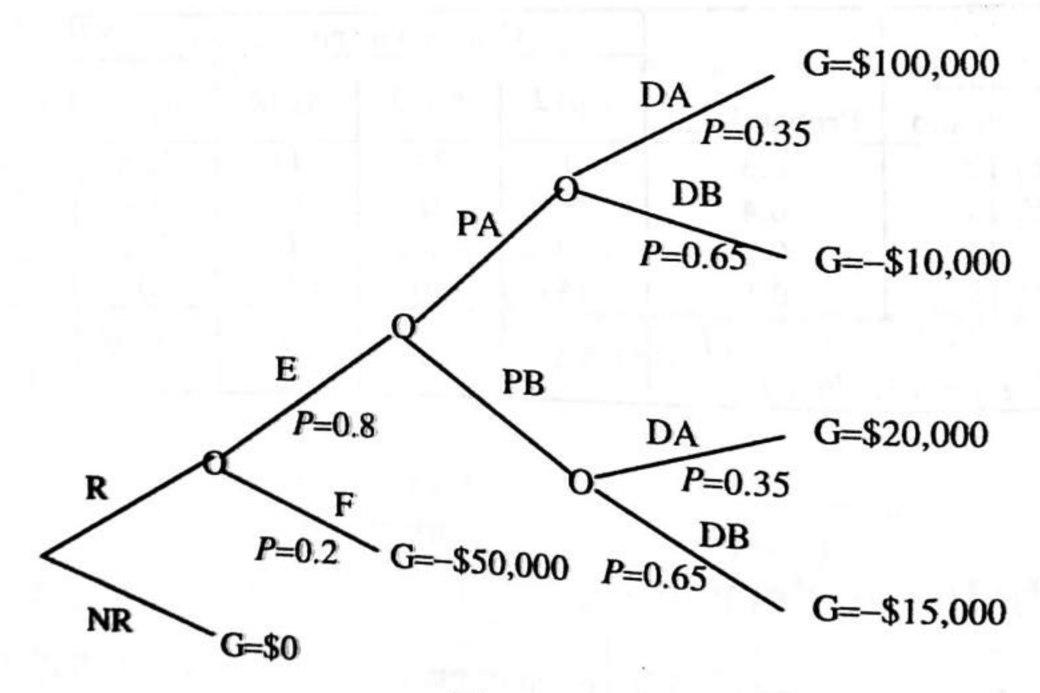


Figura 16.1

EJEMPLO 16.10
Continuando con el problema 16.9, determine si el fabricante debe realizar o no el proyecto con base a la utilidad esperada.

SOLUCION

Las utilidades esperadas en el punto de decisión de la producción son:

UE(producción alta) = 0.35×100,000 + 0.65×(−10,000) = \$28,500
UE(producción baja) = 0.35×20,000 + 0.65×(−15,000) = −\$2,750

Al comparar las dos utilidades esperadas, la mejor acción en el punto de producción es realizar una alta producción. Por lo tanto, se elimina la posibilidad de la otra acción. En el diagrama de árbol esta eliminación se indica con dos rayas en la rama correspondiente.

Regresando al punto inicial de decisión, se tienen las siguientes utilidades esperadas:

UE(Realizar) = 0.8×\$28,500 + 0.2×(−\$50,000) =\$12,800
UE(No realizar) = 0\$

Al comparar las dos utilidades esperadas, la mejor acción en el punto inicial es realizar el proyecto.

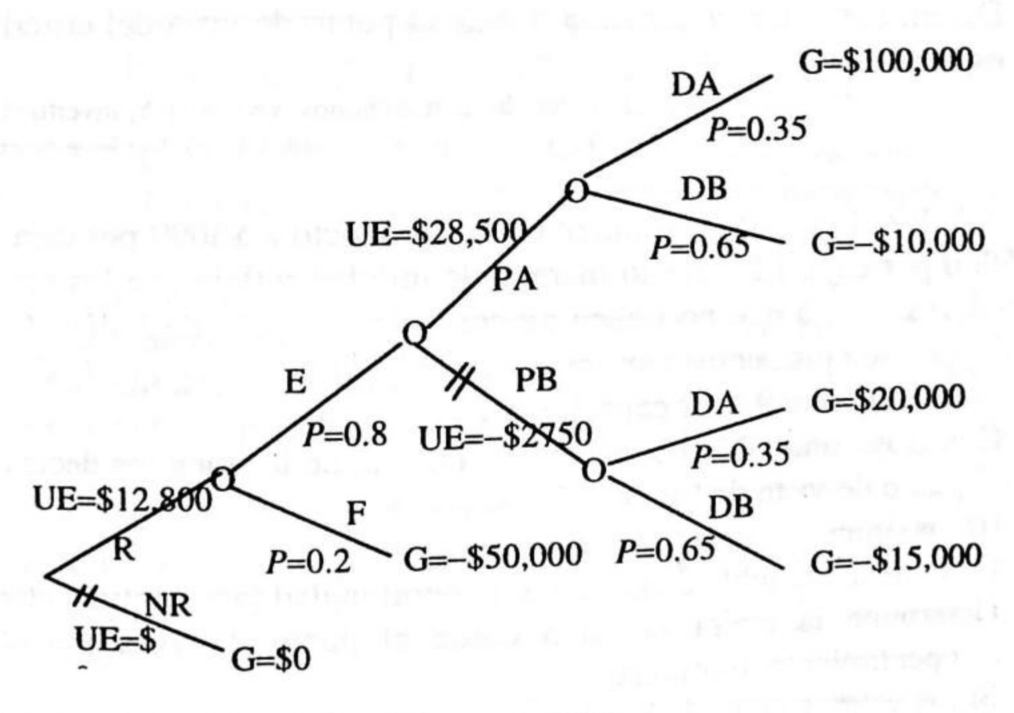


Figura 16.2

EJERCICIOS

1. Un analista de inversiones estima que existe una probabilidad de 50% de un auge en la industria química durante el primer trimestre del año y que las probabilidades de ningún cambio y recesión son iguales. Un cliente está considerando invertir $10,000,000 en un fondo de inversión que se especializa en acciones comunes de la industria química, o invertir ese dinero en bonos AAA, que producen el 28% anual. Si la industria química experimenta un auge durante el primer trimestre, el valor de las acciones de la sociedad de inversión (incluyendo dividendos) aumentará en 35% en los próximos doce meses. Si no existe cambio, el valor aumentará en 3%. Si ocurre una recesión, el valor se reducirá en 30%.
   a) Ignorando los costos por comisiones, construya una tabla de pagos para este problema de inversión.
   b) Determine las mejores decisiones desde el punto de vista de los criterios: (i) maximín, (ii) maximax.
   c) Determine la mejor decisión desde el punto de vista del criterio del arrepentimiento minimax.
   d) Determine la mejor decisión desde el punto de vista del criterio del pago esperado.

   Rp. a) (i) A_2 : Invertir en bonos AAA, (ii) A_1 invertir en un fondo de inversión b) A_2 : invertir en bonos AAA, c) A_2 : invertir en bonos AAA

2. Un vendedor al menudeo adquiere cierto producto a $3000 por caja y lo vende en $5000 por caja. El elevado margen de utilidad refleja que los productos son perecederos puesto que no tienen ningún valor después de 5 días. Con base en experiencias en productos similares, el vendedor confía en que la demanda para el artículo está entre 9 y 12 cajas inclusive.
   a) Construya una tabla de ganancias. Determine las mejores decisiones desde el punto de vista de los criterios: (i) maximin, (ii) maximax.
   b) Construya una tabla de pérdidas de oportunidad (arrepentimientos)
   c) Determine la mejor decisión desde el punto de vista del criterio del arrepentimiento minimax.
   d) Si los valores de probabilidad estimadas para las demandas de 9 a 12 cajas del producto son: 0.3, 0.4, 0.2, y 0.1 respectivamente. Determine las mejores decisiones desde el punto de vista de: (i) el criterio de probabilidad máxima, (ii) la esperanza del evento.
   e) Determine la mejor acción desde el punto de vista del criterio:
      i) del pago esperado,
      ii) de minimización de la pérdida de oportunidad esperada.

Rp. a) (i) A_1 : Ordenar 9 cajas, (ii) A_4 ordenar 12 cajas, b) A_2 : Ordenar 10 cajas, c) (i) A_2 : Ordenar 10 cajas, (ii) A_2 ordenar 10 cajas, d) (i) A_2 : ordenar 10 cajas, (ii) A_2 ordenar 10 cajas

3. Un comerciante minorista de computadoras personales estima que el margen de utilidad de cada computadora vendida es de \$ 250. Si no se venden algunas computadoras durante unos tres meses, cada computadora no vendida le produce una pérdida de \$100. Si el minorista estima que los valores de probabilidad asociados con el hecho de que venda: 4, 5, 6, ó 7 computadoras durante los 3 meses son: 0.3, 0.4, 0.2, 0.1, respectivamente.
   a) Construya la tabla de pagos para el problema.
   b) Determine la mejor acción a tomar utilizando:
      i) el criterio de pago esperado
      ii) el criterio de pérdida de oportunidad esperada.

4. Un panadero produce cierto tipo de pan francés por la noche a un costo de \$ 0.10 cada uno y lo vende al día siguiente a \$1.5 cada uno. Este pan es perecedero y se debe desechar si no se vende durante el día. De acuerdo con la experiencia, la demanda diaria (en miles de unidades) y las respectivas probabilidades son las siguientes:

Demanda	1	2	3	4
Probabilidad	0.2	0.4	0.3	0.1

   a) Construya una tabla de pagos para este problema de negocio.
   b) Determine la mejor acción utilizando el criterio del pago esperado.

5. Continuando con el problema 4:
   a) Construya una tabla de pérdidas de oportunidades condicionales
   b) Indique la acción óptima utilizando el criterio de las pérdidas de oportunidades esperadas.

6. Cierto producto se produce a S/.5 soles y se vende a S/.10 la unidad. Este producto es tal que si se produce y no se vende durante una semana, se deshecha sin costo alguno. Los registros de ventas semanales en el pasado son los siguientes:

Demanda	17	18	19	20
Semanas	100	500	300	100

   a) Construya una tabla de pagos para este problema de negocio.
   b) Determine la mejor acción utilizando el criterio del pago esperado.

7. Continuando con el problema 6
   a) Construya una tabla de las pérdidas de oportunidades condicionales
   b) Determine la acción óptima utilizando el criterio de las pérdidas de oportunidades esperadas.

8. Varias semanas antes de la temporada de vacaciones, un grupo de comerciantes debe decidir si ordena colchones o sombrillas de playa o nada, para venderlos en un centro de recreo. El éxito de las ventas depende de las condiciones del clima frío o caliente. Si la tabla de pagos es:

Clima	A: Colchones	A : Sombrillas	A : Ninguna
Demanda	\$ 100	–\$ 80	\$ 0
Semanas	– 50	150	0

   a) Determine las mejores decisiones desde el punto de vista de los criterios: i) maximin, ii) maximax.
   b) Determine la mejor decisión desde el punto de vista del criterio del arrepentimiento minimax.

Rp. a) (i) A_3 : nada, (ii) A_2 sombrilla de playa, b) A_3 : nada

9. Una tienda de comestibles compra diariamente a \$5 por unidad un producto que vende a \$10 cada uno. Cada unidad que se vende le cuesta a la tienda \$0.50 por envoltura. Debido a que el producto es perecedero, las unidades que se quedan sin vender al final del día son devueltas sin envoltura al proveedor para nuevo procesamiento y se recibe una devolución de \$2 por unidad. Se estima que los valores de probabilidad asociados con el hecho de que venda: 1, 2, 3, 4 ó 5 unidades durante el día son 0.1, 0.4, 0.3, 0.1, 0.1, respectivamente.
   a) Construya la tabla de pagos para el problema.
   b) Determinar la mejor acción a tomar utilizando:
      i) el criterio de pago esperado
      ii) el criterio de pérdida de oportunidad esperada.

Rp. a) pagos: $D = 1$: 4.5, 1.5, −1.5, −4.5, −7.5. D=2: 4.5, 9, 6, 3, 0. D =3: 4.5, 9, 13.5, 10.5, 7.5. D=4: 4.5, 9, 13.5, 18, 15. D=5: 4.5, 9, 13.5, 18, 22.5, b) Mejor decisión: $A = 3$.

10. A un fabricante se le presentó una proposición para un producto nuevo y debe decidir desarrollarlo o no. El costo del desarrollo del proyecto es \$200,000,000. La probabilidad del éxito es 0.7. Si el desarrollo no tiene éxito, se termina el proyecto. Si tiene éxito, el fabricante entonces ha de decidir si el nivel de producción ha de ser alto o bajo. Si la demanda es alta, el aumento en la utilidad,

dado un nivel alto de producción es de $700,000,000; dado un nivel bajo es de $150,000,000. Si la demanda es baja, el incremento en la utilidad, dado un nivel elevado de producción es $100,000,000; dado un nivel bajo es de $150,000,000. Todos estos incrementos en utilidades son cifras brutas (es decir antes de restar los $200,000,000 del costo de desarrollar el proyecto). Si se estima que la probabilidad de una demanda elevada es 0.40 y para la demanda alta es 0.6,

a) Elabore un árbol de decisiones para esta situación.

b) Utilizando el árbol de decisiones, determine si debe realizarse o no el proyecto.

Rp. Desarrolle el proyecto, PE(desarrollar) = $38,000

11. Un inversionista debe decidir si realiza o no un depósito de $50,000 para construir un mercado en una área residencial. Es posible que un competidor importante decida también abrir un mercado en la misma área y, por otro lado, no se sabe si esta área residencial crecerá para convertirse en un mercado grande o moderado. El inversionista estima una probabilidad de 0.50 de que el competidor establezca una tienda. Si existe competencia y el mercado es grande la ganancia neta se estima en $75,000; si el mercado es moderado habrá una perdida neta de 50,000. Si no existe competencia y el mercado es grande, la ganancia neta será de $150,000; si el mercado es moderado, habrá una ganancia neta de $50,000. Si el inversionista estima una probabilidad de 40% de que el mercado sea grande. Utilizando un diagrama de árbol de decisiones determine si el inversionista debe realizarse o no el depósito.

Rp. $45,000.

12. Un inversionista está considerando colocar un depósito de $10,000,000 para reservar una oportunidad de concesión en una nueva área residencial durante un año. Existen dos áreas de incertidumbre asociadas con esta situación de situaciones secuenciales: en primer lugar, es posible que un competidor importante de otra concesionaria decida abrir un expendio en la misma área y, por otro lado, no se sabe si esta área residencial crecerá para convertirse en un mercado moderado o grande. El inversionista estima una probabilidad de 0.50 de que el sistema competidor establezca una tienda. Por ello, lo primero que el inversionista debe decidir es si realiza el pago inicial de $10,000,000. Después que se sepa la decisión del sistema competidor, el inversionista debe decidir si ha de proceder o no a construir su propio expendio. Si existe competencia y el mercado es grande la ganancia neta para el periodo de interés se estima en $15,000,000; si el mercado es moderado habrá una perdida neta de 10,000,000. Si no existe competencia y el mercado es grande, la ganancia neta será de 30,000,000; si el mercado es moderado, habrá una ganancia de $10,000,000. Si

el inversionista estima una probabilidad de 4% de que el mercado sea grande. Utilizando un diagrama de árbol de decisiones determine si debe realizarse o no el depósito.

Rp. Realizar el depósito ($PE = \$9{,}000{,}000$).

13. Un señor posee un terreno que puede utilizarse como estacionamiento. El encargado de un garaje ofrece al dueño dos alternativas para la utilización de su terreno: Una, un contrato sin pago de alquiler por el terreno, basado en participación de los beneficios; la otra una compra directa del terreno por $50,000. Además, se ha estimado que la primera alternativa producirá una renta de $80,000 si el negocio resulta un éxito. Si fracasa, la renta será sólo de $20,000. ¿Qué probabilidad se debe ligar al éxito para que el dueño sea indiferente a las dos alternativas si debe decidir con base en los pagos esperados?.

Rp. 0.50

14. El gerente de ventas de una librería ha determinado que la **demanda** por semestre del libro Estadística Inferencial de M.Córdova es una variable aleatoria X con distribución de probabilidad

X	10	20	30	40	50	60
$P[X = x]$	0.1	0.1	0.2	0.3	0.2	0.1

Cada libro la librería compra a S/. 8 y lo vende a S/. 15. Cada libro que no venda en el semestre le produce una pérdida adicional de $1. El gerente de ventas quiere determinar cuántos libros comercializar por semestre. Usted ayúdelo contestando las siguientes preguntas.

a) ¿ Cuántos libros debería comercializar por semestre, si la decisión se basa en la demanda más probable?.

b) ¿ Cuántos libros debería comercializar por semestre, si la decisión se basa en la demanda promedio?.

c) Si la librería decide comercializar 30 libros cada semestre, ¿cuánto espera de utilidad?.

d) ¿ Cuántos libros debería comercializar por semestre, para maximizar la utilidad esperada?

Rp. a) 40 libros, b) $E(X)$=37 , c) Para 30, Utilidades respectivas:−110, 50, 210, 210, 210, 210, utilidad esperada = S/.162. , d) Para 20, $/400, para 21, $413.125, para 23, $370.625, es máxima para 21.

15. La demanda por temporada en miles metros de determinada tela que tiene una compañía textil es una variable aleatoria X con función de densidad:

$$f(x)=\begin{cases}\frac{1}{5} & \text{si } \ 0\le x\le 5\\ 0 & \text{en el resto}\end{cases}$$

Por cada metro de tela vendida se gana 3$, pero por cada metro de tela no vendida en la temporada se pierde $1. Calcular la producción que maximiza la utilidad esperada de la compañía.

Rp. X: demanda, sea K la producción Y: utilidad, $Y=3K$ si $x\ge K$, $Y=3x-1(K-x)$ si $x<K$. $E(Y)=(-2K^2/5)+3K$. $E(Y)$ es máximo si $K=(15/4)\times 1000$..

16. Una compañía alquila computadoras por periodos de tiempo de t horas, por lo cual recibe $600 por hora. El número de veces que una computadora falla en t horas es una variable aleatoria con distribución de Poisson con $\mu = 0.8t$. Si una máquina falla x veces en t horas, el costo de reparación es $\$50x^2$. ¿Cómo debería la compañía elegir t en forma tal que maximice su utilidad esperada?

Rp. Utilidad: $600t-50X^2$, $E(\text{Utilidad})=600t-50E(X^2)=600t-50(\text{Var}(X)+(E(X))^2)=600t-50(0.8t+(0.8t)^2)=600t-40t-32t^2$. Derivando respecto de t, e igualando a 0: $600-40-64t=0$, $64t=560$, $t=8.75$

17. El distribuidor de un producto de temporada tiene la política de comprar al inicio de la misma una existencia de K unidades de volumen a 10 soles por unidad. Durante la temporada vende el producto a 25 soles la unidad de volumen y al final de la temporada, si le queda un sobrante, lo debe desechar, sin costo alguno.

Se sabe que la cantidad de producto demandada X (en toneladas), al distribuidor es una variable aleatoria continua con función de densidad

$$f(x)=e^{-x}, \quad x\ge 0$$

a) Halle el nivel de producción que maximiza la utilidad esperada

b) Halle la probabilidad de que la demanda supere al nivel de producción óptimo.

Rp. $U=25X-10K$ si $X\le K$, $U=15K$ si $X\le K$, $E(U)=25\int_0^K xf(x)dx-25KP[X\le K]+15K$, $d(U)/dk=0$ da $1-e^{-K}=3/5$, $K=-\ln(2/5)=0.916$

18. El diámetro interior (en milímetros) de un tubo, es una variable aleatoria X distribuida normalmente con media μ y varianza 1. Si $10 \leq X \leq 12$, la utilidad por tubo del fabricante es de \$10. Si $X<10$, la utilidad por tubo es de $-\$3$, y si $X>12$, la utilidad es de $-\$2$.

   a) Determine la utilidad esperada por tubo del fabricante

   b) Suponga que el fabricante puede ajustar su proceso de fabricación para diferentes valores de μ. ¿En cuánto debería ajustar el fabricante el valor de μ, para que maximice su utilidad esperada?

   (Puede usar $F'(z) = f(z)$ donde $f(z) = (2\pi)^{-1/2} e^{-Z^2/2}$

Rp. a) $E(U)=12F(12-\mu)-13F(10-\mu)-2$, b) $E'(U)=0$, da $e^{22-2\mu}=12/13$.
$\mu=11-1/2\ln(12/13)=11.04$

Apéndice :

Estudio socioeconómico de estudiantes universitarios de Lima.

X_1: Sexo (1 = Mujer, 0 = Hombre)

X_2: Edad,

X_3: Origen (0 = Costa, 1 = Sierra, 2 = Selva)

X_4: Colegio (1 = Nacional, 0 = Particular)

X_5: Año de ingreso a la U

X_6: Ingreso familiar mensual

X_7: Tamaño de la familia

X_8: Gasto mensual en educación

X_9: # Hermanos estudiantes

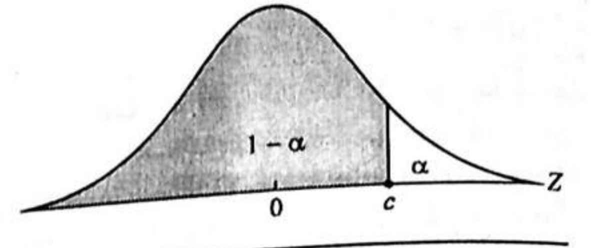
X_{10}: Casa propia (1 = Si, 0 = No)

X_{11}: Tenencia de automóvil (1 = Si, 0 = No)

#	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	1	19	2	0	1995	2300	5	1000	3	1	0
2	1	20	0	0	1995	2780	4	1200	2	1	0
3	0	23	1	1	1994	3000	5	1500	2	1	0
4	1	24	0	1	1997	4000	7	2200	4	0	1
5	0	20	0	0	1998	4500	9	3000	6	1	0
6	0	26	0	0	2000	2800	6	1100	0	1	0
7	1	19	1	0	1999	7000	8	3800	5	1	1
8	1	19	1	1	2001	5000	6	2000	1	0	1
9	0	20	0	1	2000	4700	8	3200	4	0	1
10	1	21	0	1	1998	1200	3	600	1	1	0
11	0	20	0	0	1999	3900	9	2100	2	1	0
12	1	23	0	1	1994	5200	12	3700	4	1	1
13	0	24	2	0	1997	5300	10	3400	3	0	1
14	0	24	1	0	1999	3800	7	2000	4	1	0
15	0	22	0	1	1998	2500	6	1300	2	1	0
16	1	21	1	1	1996	4200	8	3100	4	0	1
17	1	21	1	0	1997	1500	7	800	2	1	0
18	1	20	1	1	2000	8000	6	3000	2	0	1
19	1	26	0	0	2001	5200	7	1300	2	1	1
20	0	19	0	1	2002	2700	5	700	1	0	0
21	0	20	2	1	2000	3100	5	1340	1	1	0
22	0	22	1	1	2001	2450	7	1200	2	0	0
23	1	23	0	0	2001	3500	6	1100	1	1	1
24	1	24	1	1	1999	4100	8	2670	5	1	1
25	0	21	1	0	1998	2300	5	800	1	1	0
26	1	20	2	0	1997	1900	5	600	1	0	0
27	0	28	1	0	1996	1000	3	300	0	0	0
28	0	25	1	1	1997	2580	6	1250	2	0	0
29	0	22	0	0	1995	3500	8	1370	3	1	1
30	0	21	1	1	1994	3200	6	1250	0	1	0
31	1	20	0	0	1996	9000	9	3800	1	0	1
32	0	21	2	0	1999	2600	5	1500	1	1	0
33	1	22	0	0	2000	3790	8	2400	2	0	1

TABLA DE LA DISTRIBUCION NORMAL ESTANDAR

La tabla da áreas $1-\alpha$ y valores
$c = Z_{1-\alpha}$, donde, $P[Z \leq c] = 1-\alpha$,
y donde Z tiene distribución normal $N(0,1)$.

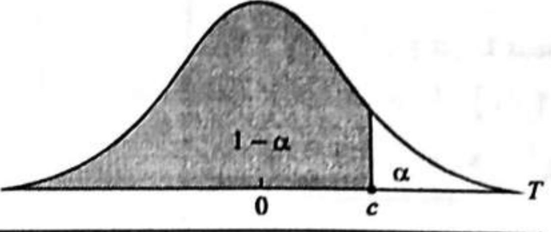


Z	Segundo decimal de Z .00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9762	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	1.000	1.000	1.000
4.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

M. Córdova Z.

TABLA DE LA DISTRIBUCION *t*–Student

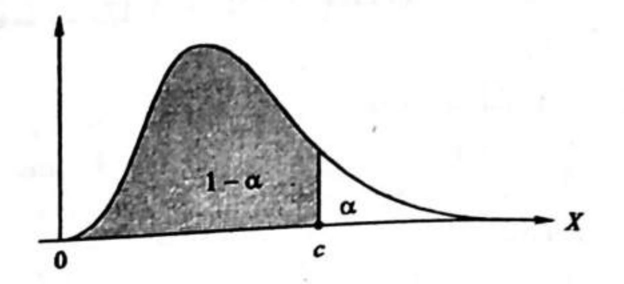
La tabla da áreas $1-\alpha$ y valores $c = t_{1-\alpha,r}$, donde, $P[T \leq c] = 1-\alpha$, y donde T tiene distribución *t*-Student con r grados de libertad.



	$1-\alpha$							
r	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

TABLA DE LA DISTRIBUCION CHI-CUADRADO

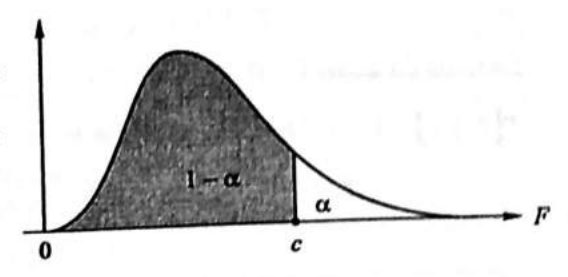
La tabla da áreas $1-\alpha$ y valores $c=\chi^2_{1-\alpha,r}$ tales que $P[X \le c]=1-\alpha$, donde X tiene distribución χ^2 con r grados de libertad.



r	1 − α									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.71	3.84	5.02	6.64	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.82	9.35	11.35	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.65	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.58	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.69	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.27	7.01	8.23	9.39	10.87	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.57	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

TABLA DE LA DISTRIBUCION DE PROBABILIDADES F

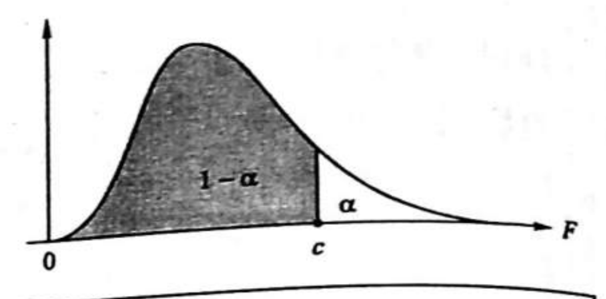
La tabla da áreas $1-\alpha$ y valores $c = F_{1-\alpha, r_1, r_2}$ tales que $P[F \le c] = 1-\alpha$. Donde r_1 y r_2 son los grados de libertad, y donde $F_{\alpha, r_2, r_1} = 1/F_{1-\alpha, r_1, r_2}$.



$1-\alpha$	r_2	r_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	120
.95	1	161	200	216	225	230	234	237	239	241	242	244	246	248	253
.975		648	800	864	900	922	937	948	957	963	969	977	985	993	1014
.95	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5
.975		38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5
.99		98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5
.995		199	199	199	199	199	199	199	199	199	199	199	199	199	199
.95	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.55
.975		17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	13.9
.99		34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.2
.995		55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.4	43.1	42.8	42.0
.95	4	6.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.66
.975		12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.31
.99		21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.6
.995		31.3	26.9	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.7	20.4	20.2	19.5
.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.40
.975		10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.07
.99		16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.11
.995		22.8	18.3	16.5	15.6	14.9	14.5	14.2	[illegible]	13.8	13.6	13.4	13.1	12.9	12.3
.95	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.70
.975		8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	4.90
.99		13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	6.97
.995		18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.2	10.0	9.81	9.59	9.00
.95	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.27
.975		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.20
.99		12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.74
.995		16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.19
.95	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	2.97
.975		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.73
.99		11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	4.95
.995		14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.06
.95	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.75
.975		7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.39
.99		10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.40
.995		13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.30

TABLA DE LA DISTRIBUCION DE PROBABILIDADES F

La tabla da áreas $1-\alpha$ y valores $c = F_{1-\alpha, r_1, r_2}$ tales que $P[F \le c] = 1-\alpha$. Donde r_1 y r_2 son los grados de libertad, y donde $F_{\alpha, r_2, r_1} = 1/F_{1-\alpha, r_1, r_2}$.



$1-\alpha$	r_2	r_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	120
.95	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.58
.975		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.14
.99		10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.00
.995		12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	4.75
.95	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.34
.975		6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.79
.99		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.45
.995		11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.01
.95	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.11
.975		6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.46
.99		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	2.96
.995		10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.37
.95	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	1.90
.975		5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.16
.99		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.52
.995		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	2.81
.95	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.68
.975		5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	1.87
.99		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.11
.995		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.30
.95	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.47
.975		5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.58
.99		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	1.73
.995		8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	1.83
.95	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.35
.975		5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.95	1.82	1.43
.99		6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.53
.995		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.61
.95	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.22
.975		5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.27
.99		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.32
.995		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.36

Tabla de la prueba de una muestra de Kolmogorov-Smirnov

Tabla de valores críticos de en la prueba de una muestra de Kolmogorov-Smirnov, donde:

$D = \text{máximo } \left| F_n(X) - S_n(X) \right|$

Si $D >$ valor crítico de la tabla, se rechaza H_0.

Tamaño de la Muestra n	Nivel de significación, α				
	0.20	0.15	0.10	0.05	0.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.21	.22	.24	.27	.32
30	.19	.20	.22	.24	.29
35	.18	.19	.21	.23	.27
36 o más	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

R Meléndez

BIBLIOGRAFIA

1. Amón Jesús. Estadística para Psicólogos. Ediciones Pirámide.
2. Calzada Benza José. Estadística General. Editorial Jurídica. Lima.
3. D'Ottone Horacio. Estadística Elemental. CCP. Chile.
4. Fausto I. Toranzos. Estadística. Kapelusz. Buenos Aires.
5. Levin Richard. Estadística para Administradores. Prentice Hall.
6. Lincoln L. Chao. Introducción a la Estadística. CECSA.
7. Lincoyán Portus Goviden. Curso Práctico de Estadística. McGraw-Hill
8. Lind, Mason y Marchal ESTADÍSTICA para Administración y Economía 3era Edición (2000), McGraw-Hill, 575 páginas.
9. Mood Graybill. Introducción a la Teoría de la Estadística. Aguilar.
10. Meyer P.L.. Probabilidad y Aplicaciones Estadísticas. Fondo Educativo Interamericano.
11. Miller y Freund. Probabilidad y Estadística Para Ingenieros. Prentice Hall.
12. Ostle Bernardo. Estadística Aplicada. Limusa.
13. Siegel Sydney. Estadística No Parámetrica. Trillas.
14. Spiegel Murray. Probabilidad y Estadística. McGraw-Hill.
15. Taro Yamane. Estadística. Editorial Harla.3era edición
16. Walpole y Myer. Probabilidad y Estadística. McGraw-Hill.
17. Ya-Lun Chou. Análisis Estadístico. Interamericana.
18. Manual del SPSS
19. Manual del ESTADISTICA